IJCSIS Vol. 12 No. 1, January 2014
ISSN 1947-5500

International Journal of Computer Science & Information Security


Cornell University
Library


Cogprints


Google scholar


.docstoc
find and share professional documents


ScientificCommons

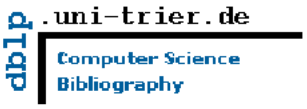

View my documents on
Scribd


BASE
Bielefeld Academic Search Engine


SCIRUS
search engine for science


SciRate.com


CiteSeerx beta


dblp .uni-trier.de
Computer Science
Bibliography


Q·Sensei


DOAJ
DIRECTORY OF
OPEN ACCESS
JOURNALS


EBSCO
HOST


ProQuest

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2014 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Managing Editor

International Journal of Computer Science and Information Security *(IJCSIS – established since May 2009), is a global venue to promote research and development results of high significance in the theory, design, implementation, analysis, and application of computing and security. As a scholarly open access peer-reviewed international journal, IJCSIS aims at providing a platform and encourages emerging scholars and academicians globally to share their professional and academic knowledge in the fields of computer science, engineering, technology and related disciplines. This journal is also particularly interested in bridging the gap between theoretical computer science and its practical applications in the real-world. Thus, papers that can provide both theoretical analysis coupled with carefully designed experiments are welcomed.*

IJCSIS archives all publications in major academic/scientific databases; abstracting/indexing, editorial board and other important information are available online on homepage. Indexed by the following International agencies and institutions: Google Scholar, Bielefeld Academic Search Engine (BASE), CiteSeerX, SCIRUS, Cornell's University Library EI, Scopus, DBLP, DOI, ProQuest, EBSCO. Google Scholar reported increased in number cited papers published in IJCSIS. IJCSIS supports the Open Access policy of distribution of published manuscripts, ensuring "free availability on the public Internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of [published] articles".

IJCSIS editorial board ensures a rigorous peer-reviewing process and consisting of international experts. IJCSIS solicits your contribution with your research papers. IJCSIS is grateful for all the insights and advice from authors & reviewers.

We look forward to your collaboration. Get in touch with us. For further questions please do not hesitate to contact us at ***ijcsiseditor@gmail.com****.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 12, No. 1, January 2014 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

IJCSIS EDITORIAL BOARD

TABLE OF CONTENTS

smoothie and easy. However, RFID also has some negative issues such as violation to the people's privacy and violation to the data protection. Even though, these security issues RFID has changed the way dealing with products and people. Thus, let's take this opportunity, improve the life, use new technology even if there is a little fair do not hesitate to take the benefit from every new technology that change our life to the best.

Keywords-component; RFID; IT technology; Networks; RFID issues; tags; SWOT Introduction

4. Paper 31121323: Two Phase K-Nearest Neighbors Approach (pp. 18-25)

Siddhartha Kumar Arjaria, Deepak Singh Tomar, Devshri Roy
Department of computer science & Engg., Maulana Azad National Institute of Technology, Bhopal, (M.P.), India

Abstract — K-nearest neighbors approach is the popular algorithm for classification. The majority of votes of neighbors of testing sample decide the class of in K-nearest neighbors approach. It only utilizes the information stored in the first few samples while it considers the remaining samples unimportant. The classification result of K-nearest neighbors approach highly depends on the single criteria, due to this classifier many times produces the wrong result. The paper presents a novel idea to deal with the classification problem in two Phases. First phase deals with the extraction of useful information from the training space regarding the occurrence behavior of each training sample in the neighbor list of other training samples. This occurring behavior decides each training sample to be part of one of the three classes namely important, unimportant, and neutral. In the second phase, On the basis of this collected information the training samples in the neighbors of testing sample are rearranged by removing the unimportant samples. Now classification decision totally omitted the unimportant training samples and considers only the important & neutral class training samples. Algorithm is designed to provide the extra weights to the important samples on the basis of its position in neighbor list, it's occurrence frequency as a neighbors of other training samples and the number of training samples of that class used for training. Performance is tested on three database seven most frequent categories of Reuters-21578, four most frequent categories of RCV1, seven most frequent categories of TDT2 corpus. Our approach outperforms K-nearest neighbors approach in terms of F1 value in almost each case.

Keywords- K-nearest neighbors approach;Two Phase KNN;Classification;

5. Paper 31121327: Developing Extracting Association Rules System from Textual Documents (pp. 26-36)

Arabi Keshk and Hany Mahgoub
Faculty of Computers and Information, Menoufia University, Shebin El-Kom, Egypt

Abstract — A new algorithm is proposed for generating association rules based on concepts and it used a data structure of hash table for the mining process. The mathematical formula of weighting schema is presented for labeling the documents automatically and its named fuzzy weighting schema. The experiments are applied on a collection of scientific documents that selected from MEDLINE for breast cancer treatments and side effects. The performance of the proposed system is compared with the previous Apriori-concept system for the execution time and the evaluation of the extracted association rules. The results show that the number of extracted association rules in the proposed system is always less than that in Apriori-concept system. Moreover, the execution time of proposed system is much better than Apriori-concept system in all cases.

Keywords- data mining; association rules; fuzzy system; apriori-concept system

6. Paper 30111336: A Novel Congestion Control Mechanism for Traffic Management in Wireless Sensor Networks (pp. 37-41)

Nasrin Azizi, Solmaz Abdollahi Zad
Department of Computer, College of Computer, Sardroud Branch, Islamic Azad University, Sardroud, Iran

Abstract - Due to the nature of wireless sensor networks the higher amount of traffic is observed when the monitored event takes place. Exactly at this instance, there is a higher probability of congestion appearance in the network. Congestion can cause missing packets, low energy efficiency, and long delay. Moreover, some applications, e.g. multimedia and image, need to transmit large volumes of data concurrently from several sensors. These applications have different delay and QoS requirements. Congestion problem is more urgent in such applications. Therefore congestion in WSNs needs to be controlled for high energy-efficiency, to prolong system lifetime, improve fairness, and improve quality of service in terms of throughput and packet loss ratio along with the packet delay. To achieve this objective, a novel congestion control protocol for traffic management is proposed in this paper. Proposed protocol can control congestion in the node and adjusts every upstream traffic rate with its node dynamic priority to mitigate congestion. Proposed protocol can broadcast traffic on the entire network fairly. Simulation results show that the performance of proposed protocol is more efficient than previous algorithms in terms of throughput.

Keywords-wireless sensor network; congestion mitigation; traffic distribution; throughput

7. Paper 31121308: Email Security Using Clustering Algorithms (pp. 42-48)

Tarushi Sharma, M-Tech(Information Technology), CGC Landran Mohali, Punjab, India

Abstract — Recent use of email analysis and data mining of email contents has proven to be useful in some sensitive places like national security agency to detect threats and fraud determination from terrorists. Moreover, it has been proved to be helpful for decision making, future team co-ordination, fraud detection and tracing the behavior of an employee. Using different clustering algorithms, we can find out similar patterns in emails for fraud detection. In this paper, we demonstrate how the popular k-means clustering algorithm can be profitably modified to make use of this information.

Index Terms — Data mining, Email mining, Clustering algorithm, K-mean clustering algorithm, National Security Agency and Fraud Detection

8. Paper 31121309: Email Security Using Weka Tool Results Of K-Mean Clustering Algorithm (pp. 49-54)

Tarushi Sharma, M-Tech(Information Technology), CGC Landran Mohali, Punjab,India

Abstract — Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Weka is a type of data mining tools. It is contain the many machine leaning algorithms. It provides the facility to classify our data through various algorithms. In this paper we are studying the various clustering algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters. The K-means algorithm is a popular data-clustering algorithm. However, one of its drawbacks is the requirement for the number of clusters, K, to be specified before the algorithm is applied. This paper first reviews existing methods for selecting the number of clusters for the algorithm. Our main aim to show the comparison of the different samples of data like we are using different Emails with similar text through different clustering algorithms of Weka and find out which parameter of Weka tool is effective for the users for data mining or email mining.

Keywords— Data mining algorithms, Weka tools, K-means algorithms, Clustering methods E-mail mining etc

An Enhanced Multi-Pager Environment Support for Second Generation Microkernels

Yauhen Klimiankou

Department of Information Technologies Software, Belarusian State University of Informatics and Radioelectronics
Minsk 220 113, Belarus
klimenkov@bsuir.by | Evgeny.Klimenkov@gmail.com

***Abstract*—The main objective of this paper is to present a mechanism of enhanced paging support for the second generation microkernels in the form of explicit support of multi-pager environment for the tasks running in the system. Proposed mechanism is based on the intra-kernel high granularity pagers assignments per virtual address space, which allow efficient and simple dispatching of page faults to the appropriate pagers. The paging is one of the major features of the virtual memory, which is extensively used by advanced operating systems to provide an illusion of elastic memory. Original and present second generation microkernels provide only limited, inflexible and unnatural support for paging. Furthermore, facilities provided by current solutions for multi-pager support on the runtime level introduce an overhead in terms of mode switches and thread context switches which can be significantly reduced. Limited paging support limits the attractiveness of the second generation microkernel based systems use in real-life applications, in which processes usually have concurrent servicing of multiple paging servers. The purpose of this paper is to present a facilities for the efficient and flexible support of multi-pager environments for the second generation microkernels. A comparison of the proposed solution to the present architecture L4 + L4Re has been made and overhead of the page fault handling critical path has been evaluated. Proposed solution is simple enough and provides a natural and flexible support of multi-pager environments for second generation microkernels in efficient way. It introduces a third less overhead in terms of the mode switches and thread context switches in comparison to the present L4 + L4Re solution implemented in the Fiasco.OC.**

***Index Terms*—memory management, page fault, second generation microkernel, multi-pager environment**

I. INTRODUCTION

This paper describes the design of the fine-grained multi-pager environment support facilities for the second generation microkernels that allows processes running in the system to be serviced by multiple pager servers concurrently and in efficient way. The proposed approach describes modifications of the virtual memory management subsystem of the second generation microkernels.

The user mode page fault handling was originally proposed by the Mach project [1] [2]. The general idea of the proposed approach is to allow the page fault handling servers to be running as a separate user mode processes. Safe and efficient memory management is a fundamental requirement for a microkernel. Due to this, for example, substantial memory overhead imposed by originary recursive address space construction can be considered as an enough drawback to reject this memory management approach [3]. Traditional approach for paging support in the L4 like microkernels family is limited and not efficient for multi-pager environments, which are ordinary for the advanced real-world systems and applications. Insufficient support of such environments limits the attractiveness of second generation microkernel based operating systems for the real system implementations. Furthermore this insufficient support of paging looks inadequate to the actual state of facts, because experience gained from the ubiquitous monolithic kernels shows that the typical application are likely serviced by multiple memory management subsystems concurrently. For example, the typical processes in Windows [4] and Linux [5] environment concurrently get the next services: automatic stack expanding/reducing, dynamically loadable modules management, anonymous memory management, unswappable memory management, shared memory management, file mappings to memory etc. Furthermore some specific applications can wish to use special purpose memory management facilities along with ordinary memory management services. For example they can wish to use SMARTMAP-like [6] memory management for performance benefits or InkTag [7] and Gateway [8] -like memory management features to achieve additional security and reliability guarantees.

Nevertheless of the extensive use of the multi-pager environments in typical applications present second generation microkernels have a limited support of it. The original approach taken by the L4 microkernels family [9] is an optional assignment of exactly one pager for each task running in the system. This pager is responsible for handling of all page faults generated by the tasks to which it is assigned as a pager. The same approach is kept in the descendant kernels like [10]. Present version of the L4 (Fiasco.OC) [11] provides a tricky support of multi-pager environments through introduction of the additional level of indirection - region mapper. This solution has been done on the level of runtime environment system L4Re [12] developed specially for the L4 microkernel in the Technical University of Dresden. In other words, this solution is an attempt to overcome limitations imposed by the single-pager kernel design on the level of runtime environment instead of changing kernel itself, despite

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the fact that the kernel is a natural location for multi-pager environment support and that this support can be implemented in the kernel in efficient way and only with negligible violation of the minimality principle.

Region mapper is an additional layer of indirection in page fault handling introduced by L4Re for providing a multi-pager environment for applications running in the context of this runtime environment. According to this solution, each process running in L4Re environment has a special thread running inside it, which plays a role of pager for all other threads running in the same process (threads that share the same virtual address space). This special thread of L4Re-based process is called a region mapper. Its main responsibility is to manage a virtual address space layout through virtual memory management and page fault handling. Region mapper do this by managing a special table which tracks which region of the virtual address space is serviced by which memory manager. By using this table region mapper is able to route the page faults generated by process threads to the appropriate external memory manager server. As a result, the typical page fault handling process goes through next steps:

1) Page fault is generated by a thread.
2) CPU interrupts the faulted thread and gives control to the L4 microkernel.
3) L4 looks at the faulted thread to identify, which thread is a pager of it (it is a region mapper for L4Re based processes).
4) L4 suspends the faulted thread and sends page fault message for the region mapper.
5) The region mapper looks into the table by using the fault address to find out a thread which is responsible for the faulting address as a real pager.
6) The region mapper reflects the page fault message to the real pager.
7) The real pager takes an actual actions to resolve the page fault reason and restart the faulted thread.

Region manager relies to the generic abstraction of memory mapping which is called dataspace, that was initially introduced as a part of SawMill VM framework [13]. Dataspace is a generic source of resources capable to be mapped as a continuous memory region to the virtual address space like anonymous memory region, memory mapped file or device, etc., and provides only generic memory management functionality. Dataspace is a capability-protected interface implemented by L4Re but its actual implementation is provided by external thread called dataspace manager, which is in charge of the dataspace layout and its content. As a result multiple dataspaces with different implementations managed by different servers can coexist in the system concurrently. Besides handling page faults generated by threads attached to it, region mapper is also responsible for maintaining layout of the virtual address space which it services. That means that it is responsible for inserting and removing of the dataspaces to/from virtual address space. And due to this it is capable to add/remove appropriate entries to/from the mapping table mentioned above to maintain it in actual and consistent state. The actual virtual address space region represented by dataspace is populated by pages through page faults reflected by dataspace manager.

The described existing mechanism for providing the multi-pager environment in the L4 family of second generation microkernels is complex and inefficient, because it involves multiple context switches for such typical tasks as page fault handling. We would like to propose more natural, simple and efficient way of multi-pager environment support for the second generation microkernels, which moves this support implementation from the runtime environment layer into the kernel. Proposed approach introduces less processor time overhead and memory overhead with only negligible violation of the minimality principle, which can be advocated by the same arguments which are applied for intra-kernel scheduling policies implementation.

II. VIRTUAL MEMORY AND PAGE FAULT HANDLING

Emergence of the virtual memory technology made a great impact on the whole future computer systems development. The two most significant ideas behind virtual memory are:

1) Arbitrary mappings between hardware memory layout and memory layout observed by applications.
2) Transparent changing of mappings between hardware memory and memory layout observed by applications.

Both ideas rely to the explicit support of the virtual memory by underlaying CPU architecture. Significance of the virtual memory technology can be stressed by the fact that CPU uses specially dedicated block called MMU for its support. In the following discussion we will focus on the second major idea of the virtual memory which is widespread called paging.

From the paging point of view, typical system can be split into two domains: memory resource providers and memory resource consumers, relationships between which are mediated by CPU. This mediation comes in two forms: present flag in page table entry and page fault exception. The first one allows to mark virtual memory pages as stubs, which haven't any actual resources assigned. And the second one implements a way according to which the memory resource manager can be notified about attempt to access stub virtual memory page. Both this features together provide a channel of implicit communication between memory resource provider and memory resource consumer, which allows transparent dynamic memory management, which memory consumer don't need to take care about. Memory manager is able to silently get back memory allocated to consumer earlier or allocate some additional memory to it. In the same time it can be silently called by consumer in the case when it requires memory that was got back by memory manager before.

Such reasoning and understanding of the paging in context of the virtual memory lead us to a number of conceptual conclusions about paging nature.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. There is only one manager per unit of physical address space represented resources

When multiple memory managers can coexist in the system concurrently, each of them must manage its own resources, and each resource unit in the system must be managed by exactly one memory manager. The hierarchical memory managers chain implemented in L4 microkernels in which all chain entries manage the same region of memory is unnatural, overcomplicated. Furthermore the fact that the same thread can play role of both memory manager and memory consumer for the same memory unit completely violates original virtual memory concept.

B. Relationships between memory consumer and memory manager goes through a virtual address space region which consumer trust to manage to a specific manager

Primary communication channel between memory manager and memory consumer in virtual memory system is an implicit communication channel going through CPU with protocol which allows manager to silently give and return memory to/from consumer virtual address space and allows consumer to silently request resources through page fault exceptions. Memory manager is trusted entity for the memory consumer by default, because it preserves access to all resources which it provides to the consumer. But besides trust provided to the manager in regard to access to the data stored in memory, consumer must provide it a trust of virtual address space management. The only consumer responsibility is to choose to which memory manager it trusts and which region of its virtual address space.

C. Memory manager and page fault handler is a single entity

Virtual memory model assumes that the actions that are taken in reply to the page fault exception is targeted to provide resources requested by exception trigger and restart the trigger thread execution. Due to the fact that the resource providing is a responsibility of memory manager, page fault handling is its natural responsibility too. There is no big sense to distinguish memory manager and page fault handler as two different entities.

D. Multi-pager environment is natural for advanced operating systems

During long history of the virtual memory based operating systems a lot of ways of virtual memory usage have been demonstrated. Examples of these ways includes swappable and unswappable memory allocation, memory-mapped files implementation, IO devices access management, security and process isolation, shared memory management, intelligent DLL management etc. Multi-pager environment is commonly supported in the widespread industrial OS like Windows and Linux and this support is extensively used by advanced applications.

III. DESIGN AND IMPLEMENTATION OF MULTI-PAGER ENVIRONMENT SUPPORT

A. Virtual address space management

Proposed model of multi-pager environment support introduces a fundamental abstraction - user space region. User space region is a fixed size continuous part of user part of virtual address space. User space part of virtual address space is split into a fixed number of regions, each of which contains a fixed number of pages. In our experimental system we have 1020 regions per virtual address space and 1024 pages per region. Due to this our model shown in figure 1 resembles memory management architecture of Intel x86 [14], where region can be considered as an counterpart of the directory and represents 4 Mb window of virtual address space.

User space region is a fundamental unit of virtual address space management granularity. Each user space region can be managed by independent manager. Region manager have rights to map and unmap resources owned by it into any place in the managed region without any restrictions. As a result it can do this completely transparently to any thread running in the context of virtual address space witch contains that region.

Region manager plays both roles: memory manager and pager for the regions assigned to it. Due to this on the one hand all page faults occurred in the region are transparently routed for handling to the region manager assigned to it, and on the other hand that region manager can transparently reply to the page fault by mapping resources requested by it into the appropriate place of the region affected.

Threads running the context of virtual address space are responsible only for assigning region managers for each particular region of its virtual address space. By assigning the manager for the region thread provides to it trust of this region management and is unable to control particular mappings and unmapping operation. Due to the fact there are multiple regions in the same virtual address space and managers are assigned to them independently, in result proposed model represents a natural multi-pager environment with good enough management granularity.

Implementation of the proposed model introduces memory overhead in 4Kb per virtual address space. Kernel incorporates regions table into the virtual address space abstraction implementation. Region table is implemented as a memory page which contains an array of thread ids of region managers. To each user space region with sequential number N corresponds the regions table entry with the same sequential number. Regions table itself is mapped into the kernel part of the virtual address space which is located on a fixed address. As a result during page fault kernel can easily find the region table itself and identify the region manager responsible for the region to which fault address belongs, and to which kernel will send a page fault notification message. User space region id corresponding to a virtual address belonging to a user space part of virtual address space can be easily found using the next formula:

$$RID = (A_v - B_{us})/RS, \quad (1)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

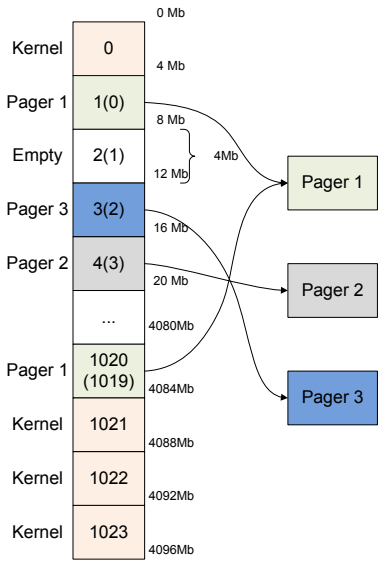


Fig. 1. Page Faults Dispatching

where RID is a region sequential number, A_v is a virtual address to which this RID corresponds, B_{us} is a base address of user space part of virtual address space and finally RS is a size of region. Note that the B_{us} and RS is a constants defined by the system design. Note also, that if RS is represented by value which is equal to power of two (which is a case of our implementation) the costly division operation can be replaced by cheap bit shift operation.

B. Page fault handling

Exceptions are a natural class of system events about which kernel must take care. In accordance with spirit of second generation microkernel design the actual work of the exception handling must be pushed out into user space and kernel must only dispatch the exception handling activities provided by external user mode servers. Due to this our experimental kernel doesn't distinguish exceptions of different types and handles them in the uniform way.

Despite the fact, that in proposed model kernel doesn't distinguish different types of exceptions and deals with all them in uniform way, the page fault exception is considered by it as a very special type of exceptions as shown in figure 2. For page faults kernel provides an additional zero level of handling and skips all other levels of handling in the case of success on that zero level.

At zero-level kernel next distinguishes two types of page faults: pure page faults and general protection page faults. The first ones are faults for addresses belonging to the user part of virtual address space and generally eligible from the protection point of view. The second ones are an faults for addresses outside of the user part of virtual address space. The kernel takes special handling only for the pure page faults and consider the another page faults as a general protection faults which are an example of the generic exceptions that must be handled in the ordinary way. By this kernel can

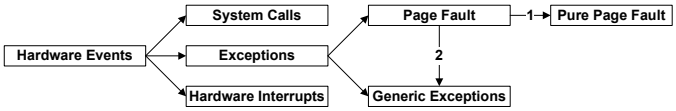


Fig. 2. Page fault classification

separate out faults that clearly aren't related to the paging and represent a clear protection violation attempt like null pointer dereferencing or attempt to access kernel code or data.

Not all pure page faults are serviced by appropriate pagers. There are two exceptional cases:

1) Faulted virtual address belongs to the virtual address space region which has not pager assigned.
2) Faulted virtual address belongs to the virtual address space region which has pager assigned but the assigned pager didn't accept the servicing of that region.

The first exceptional case is a result of multi-pager environment support. There is no single pager assigned to the thread, which is responsible for handling all page faults triggered by this thread despite the nature of page fault. Instead there is single virtual address space split into multiple regions, each of which can have pager assigned. As a result the virtual address space can be sparsely populated space, some regions of which are assigned to the pagers, and the rest have not any pagers assigned. Page fault triggered in reply to attempt to access the second ones are considered as a general protection faults.

The relationships between memory provider task (pager) and memory user task are based on the contract. To establish this relationships agreements of both sides must be received by the kernel. Memory user task provides this agreement by explicitly assigning the specified pager task to a particular virtual address space region. Memory provider task provides this agreement implicitly by taking memory management action on the region assigned to it. Furthermore pager can revoke its agreement by removing last piece of memory from the particular region with REVOKE_AGREEMENT flag specified in the system call. As a result pager task can protect itself from the malicious or misbehaving memory user task which too frequently generates page faults and by this performs DoS attack on the pager. But from the other hand this feature introduces the second exceptional case on which the page fault is generated on the region which isn't accepted by the pager. This case is also considered by the kernel as a general protection page fault.

The last check which can take place before invoking an appropriate pager is checking of the actual not presence of the appropriate memory page. On the x86 platform it can be done by checking PAGE PRESENT flag in the page table entry denoting page on which page fault occurred. This last check introduces negligible overhead, because according to our memory management subsystem implementation, kernel always reads page table entries during page fault handling to get 31-bit pager defined marker from the page table entry corresponding to the faulted page. This check can have sense,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

because there are multiple threads running in the same virtual address space allowed, which can potentially fault in the same page in very little period of time. As a result the situations are possible, in which the faulted page can be restored by the pager between page fault and actual pager invoking.

Let's consider the case illustrated on Fig. 3. In this case, there are two threads A and B concurrently running in the same virtual address space. Each of them made page fault on the same page N sequentially one after another. But thread A and thread B achieved different handling from the kernel side. Kernel notified pager that thread A triggered page fault on page N via message and blocks thread A execution until pager will have page N restored. Then when pager got a CPU time it restored mapping of the page N and unblocked thread A, allowing it future execution.

The case of thread B differs from the case of thread A in the fact, that between actual page fault generation and the end of first phase of page fault handling, pager has scheduled for CPU time and already restored mapping of the page N. As a result, kernel can simply and safely return control back to the thread B without its blocking/unblocking and additional pager involving.

IV. CASE STUDY, PRELIMINARY EVALUATION AND DISCUSSION

Lets consider three page fault handling schemes used in multi-pager environment supporting operating systems: monolithic kernel approach, Fiasco.OC approach and finally proposed approach (Fig. 4). For each case we consider a user mode thread that triggers a page fault exception which must be handled by paging server to allow the faulted thread to continue its execution. We consider a general path of the page fault handling without taking into the account the performance penalty introduced by paging server, cost of transition between kernel mode and user mode and cost of IPC.

Originally, the first proposed page fault handling cycle was proposed for monolithic kernel design. According to this approach all the page fault handling activities are performed in the kernel. Kernel is the only paging server of the system and can be considered as a tightly integrated set of subsystems, which includes multiple paging modules. Due to this we can consider monolithic kernel as a multi-pager environment provider. Page fault handling cycle in the environment of monolithic operating system includes two crossing of the kernel mode/user mode boundary: one is triggered by page fault exception and the second one is to return the control flow to the faulted thread. All page fault resolution actions are performed in the kernel without involving another threads. This page fault handling scheme is the most efficient, but not applicable for the second generation microkernel design.

The key design principle used in the first second generation microkernel L4 was a minimality principle, in accordance to which as much as possible functionality was pushed out from the kernel. Page fault handling is a part of functionality that was removed from the kernel. Instead of full-featured page fault handling kernel only dispatches page fault exceptions generated by CPU through message passing to the dedicated page fault handling server thread. This thread called pager is explicitly assigned to the threads running in the system on the one pager per thread basis. As a result, pager is wired not to virtual address space but to thread and it is responsible for handling all page faults generated by the thread to which it was assigned. Page fault handling cycle of the L4 is similar to the same in Mach and includes four crossings of the kernel mode/user mode boundaries and two thread context switches. But as you can see there is no multi-pager environment support implemented in the kernel.

The researchers from TUD noted multi-paging importance and proposed to implement its support on the level of runtime. They implemented this support in L4Re runtime, which creates a special pager thread per process and assigns it as a pager for each thread running in the same process. This pager thread maintains the database of mappings between virtual address space regions and pagers assigned to it. Page fault handling cycle in this case includes 6 crossings of kernel mode/user mode transitions and 3 thread context switches. At the first step CPU switches from user to kernel mode in reply to page fault exception triggered by running thread A. Kernel in reply suspends thread A and sends page fault message to the L4Re pager task assigned to task A. Pager in its turn consults with mapping database and figures out which task is responsible for resolution of page fault. Pager resends (reflects) the initial page fault message to the actual pager identified on the previous step. After this the actual pager finally can perform the actions for actual servicing of the page fault of task A.

TABLE I
COMPARISON OF PAGE FAULT HANDLING IN DIFFERENT MULTI-PAGER ENVIRONMENTS

Architecture	Mode switch count	Context switch count
Monolithic kernel	2	0
Proposed approach Single-paging L4 Mach	4	2
L4 Microkernel + L4Re (Fiasco.OC)	6	3

Proposed approach preserves the same page fault handling cycle as an original L4 kernel, but with natural intra-kernel support of multi-pager environment. On the other hand it is similar to the simplified page fault handling scheme of the Mach microkernel acceptable for the second generation microkernels. Page fault triggered by the thread causes transition from the user mode to kernel mode where kernel at the final step of its dispatching identifies the pager thread assigned to the region of address space which the faulted address belongs to. Then similar to other exception handling it suspends the faulted thread and sends page fault description message to the pager identified. After resolution of the page fault, pager notifies the kernel about resolution results and by this resumes faulted thread. Page fault handling cycle is accomplished by

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

passing control back to the resumed thread. As a result four transitions between kernel and user mode and two thread context switches are required by proposed approach for page fault handling cycle. Results are summarized in table I.

Proposed approach allows to reduce overhead of the page fault handling in terms of mode switch and thread context switch by 33.3% while preserving multi-pager environment support. In the same time it introduces only a very little additional code complexity and incurs only 4Kb of memory overhead per virtual address space. But note that L4 + L4Re approach preserves similar per address space memory overhead but on the runtime level, because code of the L4Re task and mapping database is enforced to be located on the unswappable memory, as it must eliminate page faults which can be triggered by L4Re pager thread itself.

In fact the proposed approach can be criticized from the point of view of minimality principle. But it can be advocated by the same arguments which was used for the advocation of the intra-kernel scheduling. Indeed, microkernel looks like a natural location for the multi-pager environment support. Additional code complexity is negligible and can be measured by only a few hundreds bytes of code. Unfortunately we can't provide an exact number of additional microkernel footprint bytes because the prototype of the proposed approach has been implemented as a part of written from the scratch kernel instead of changing the original L4 microkernel. But note also that despite the fact that it introduces memory overhead by one memory page per virtual address space the overall memory overhead of the system is reduced. In contrast to the L4Re approach, there is no requirements for additional region mapper task per virtual address space and resources used by it.

Cost of the transition between kernel and user modes, intra-kernel exception dispatching, thread context switch and IPC are main contributors to the page fault handling overhead. Proposed approach adds only negligible overhead in less than dozen of simple processor instructions to the original single-pager L4 page fault handling cycle critical path. This additional overhead is much smaller than the cost of the transition between kernel/user modes or thread context switches which usually takes more then hundred processor cycles. Unfortunately fair comparison of the page fault handling cost of original L4, Fiasco.OC and proposed approach is hard to take in our current environment, because in contrast to the L4 microkernels our research kernel relies on the asynchronous IPC (reasons behind this design choice are out of the scope of this paper). But we believe that the analytical comparison and discussion of the proposed approach outlined in this paper is clear and sufficient to highlight benefits of the proposed solution.

In general by this paper we wanted to advocate return of the multi-pager environment support into kernel. This can be considered as a step back to the Mach design, but with preserving the general second generation design principles and choices, and with entire simplification of the mechanisms used in accordance with minimality principle.

V. CONCLUSION AND FUTURE WORKS

The new way of the multi-pager environment support in the context of second generation microkernel based operating systems is proposed and described. It is showed that the demonstrated mechanism is superior because it introduces less overhead through reduction of number of mode switches and thread context switches performed during page fault handling cycle, provides more simple design and more flexible and natural environment for the system building. Despite the fact that this way of multi-pager support introduces some additional code complexity, this complexity is very small and can be advocated by the same arguments used for kernel-level scheduling advocation. Future work can be done on the base of this approach to explore other aspects of memory management in the context of second generation microkernels and designs of the full-featured multi-pager environment that can be built in the user mode using the proposed multi-pager support in microkernel.

REFERENCES

[1] M. Accetta, R. Baron, W. Bolosky, D. Golub, R. Rashid, A. Tevanian, and M. Young, "Mach: A new kernel foundation for unix development," 1986, pp. 93–112.

[2] S. Sechrest and Y. Park, "User-level physical memory management for mach." in *USENIX MACH Symposium*. USENIX, 1991, pp. 189–200.

[3] K. Elphinstone and G. Heiser, "From l3 to sel4 – what have we learnt in 20 years of l4 microkernels?" in *ACM SIGOPS Symposium on Operating Systems Principles (SOSP)*, Farmington, PA, USA, November 2013, pp. 133–150.

[4] M. Russinovich and D. A. Solomon, *Windows Internals: Including Windows Server 2008 and Windows Vista, Fifth Edition*, 5th ed. Microsoft Press, 2009.

[5] D. Bovet and M. Cesati, *Understanding The Linux Kernel*. Oreilly & Associates Inc, 2005.

[6] R. Brightwell, K. Pedretti, and T. Hudson, "Smartmap: Operating system support for efficient data sharing among processes on a multi-core processor," in *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, ser. SC '08. Piscataway, NJ, USA: IEEE Press, 2008, pp. 25:1–25:12.

[7] O. S. Hofmann, A. M. Dunn, S. Kim, M. Z. Lee, and E. Witchel, "InkTag: Secure applications on an untrusted operating system," in *Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, March 2013.

[8] A. Srivastava and J. T. Giffin, "Efficient monitoring of untrusted kernel-mode execution." in *NDSS*. The Internet Society, 2011.

[9] J. Liedtke, "On microkernel construction," in *Proceedings of the 15th ACM Symposium on Operating System Principles (SOSP-15)*, Copper Mountain Resort, CO, 1995.

[10] G. Klein, K. Elphinstone, G. Heiser, J. Andronick, D. Cock, P. Derrin, D. Elkaduwe, K. Engelhardt, R. Kolanski, M. Norrish, T. Sewell, H. Tuch, and S. Winwood, "sel4: Formal verification of an os kernel," in *Proceedings of the ACM SIGOPS 22Nd Symposium on Operating Systems Principles*, ser. SOSP '09. New York, NY, USA: ACM, 2009, pp. 207–220.

[11] The fiasco microkernel - overview. [Online]. Available: http://os.inf.tu-dresden.de/fiasco/

[12] L4re – the l4 runtime environment. [Online]. Available: http://os.inf.tu-dresden.de/L4Re/

[13] M. Aron, Y. Park, T. Jaeger, J. Liedtke, K. Elphinstone, and L. Deller, "The SawMill framework for VM diversity," in *Proceedings of the 6th Australasian Computer Systems Architecture Conference*. Gold Coast, Australia: IEEE CS Press, jan 2001, pp. 3–10.

[14] Intel Corporation, *Intel Architecture Software Developers Manual Volume 3: System Programming*. Santa Clara, CA, USA: Intel Corporation, 1999.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

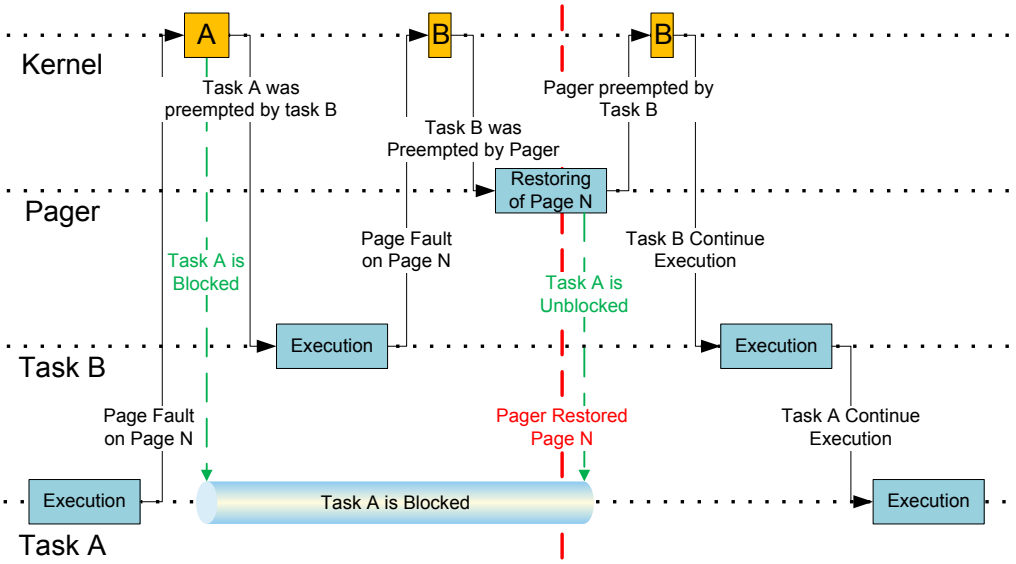


Fig. 3. Concurrent Page Fault Handling Case

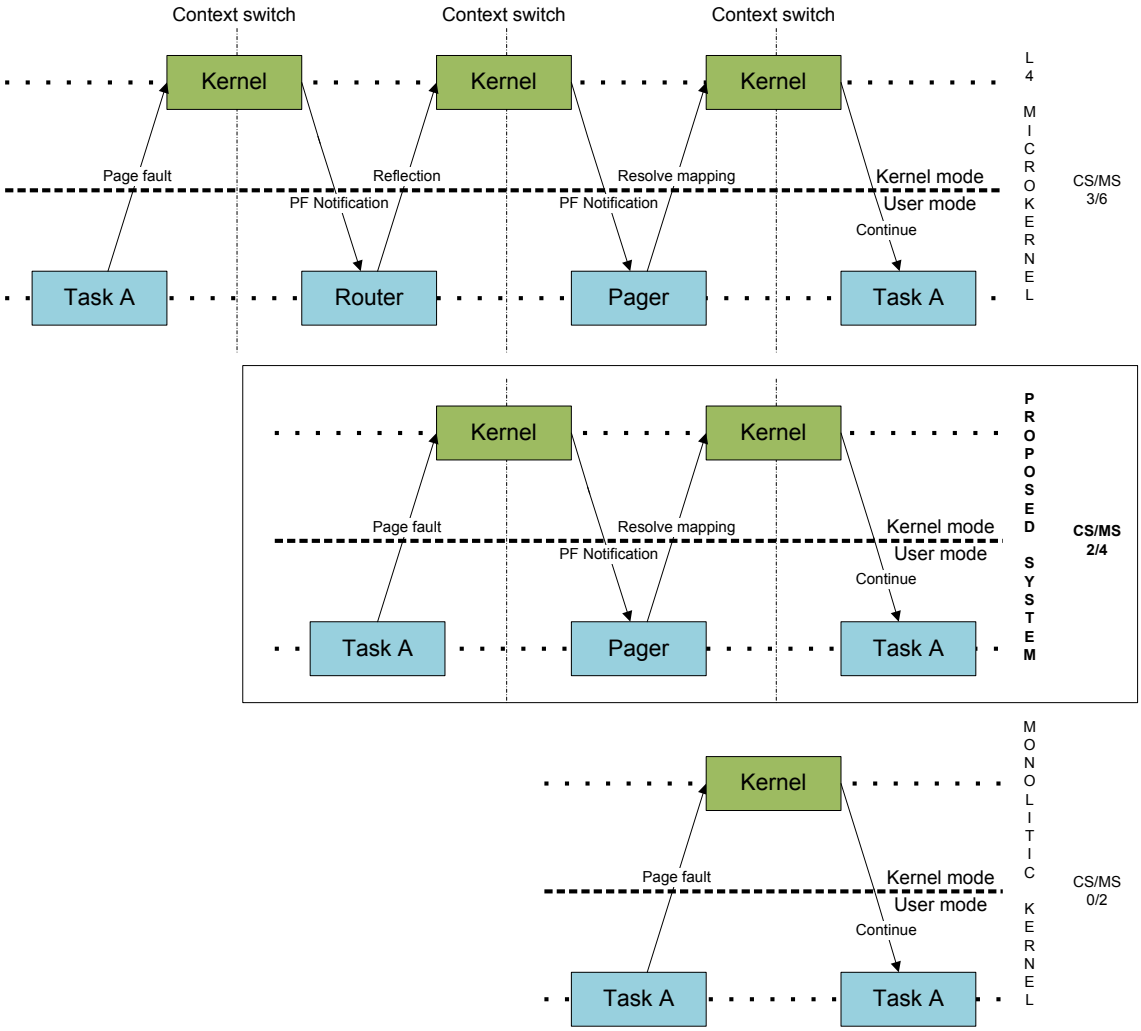


Fig. 4. Page Fault Handling Schemes

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Model performance indicators ERP systems

setare yaghubi
student, Department of computer
Zanjan Branch, Islamic Azad University
Zanjan, Iran
setareyaghubi@yahoo.com

Nasser modiri
Assoc. Prof, Department of computer
Zanjan Branch, Islamic Azad University
Zanjan, Iran
nassermodiri@yahoo.com

Masoud Rafighi
PHD Student, Department of computer
Qom University
Qom, Iran
Masoud_r62@yahoo.com

***Abstract*- Implementation process ERP is complex and expensive process. Typically always be faced with many failures. Successfully implemented in an organization has many challenges. Organizations in the deployment and success of the system depends on several factors.One of the key factors in the successful deployment of systems methodology is the implementation process. Methodology has several indicators for successful implementation of ERP systems, we have examined. And indicators for each of the methodologies have identified. The proposed method is also an important indicator of the success of security controls and indicators to be monitored and controlled.**

Keywords: Methodologies Implementation, critical success factors, ERP, AIM, ASAP, Signature.

I. Introduction

In today's dynamic and unpredictable business environment, companies face the tremendous challenge of expanding markets and rising customer expectations. This compels them to lower total costs in the entire supply chain, shorten throughput times, reduce inventories, expand product choice, provide more reliable delivery dates and better customer service, improve quality, and efficiently coordinate globe demand, supply and production [1].

The acronym ERP (Enterprise Resource Planning) was first used in 1990 by firm Gartner Group in connection to MRP (Material Requirements Planning, later Manufacturing Resource Planning) and CIM (Computer-Integrated Manufacturing) and since then it has become widely used for streamlining not only manufacturing process but all other business processes. ERP packages integrate solutions for different directions such as accounting, contracts, payroll, maintenance and human resources management, attempting to provide technology solutions for all core functions of an enterprise, regardless of its specific business flow, such as non-manufacturing businesses, non-profit organizations and governments [2].

Deploying an ERP project are including very complexity and risks. Deployment that would painlessly to replace all current software systems of Customer and all users do not prevent new systems, if not impossible, is very difficult. Because it is faced with very problems such as don't ready information, user countering with change, impossible convert operation and some case reenter information. For countering with these problems need to deployment methodology. The methodology should be an important indicator of the success of ERP systems in organizations.

One of the main points that exist in implementing large information systems, using a methodology developed and implemented the system. This is very important in case of ERP systems due to the special nature and the impress all the processes and activities of the organization.

Section 2 in this paper we study important factors in the successful deployment of ERP systems. Section 3 introduces the indicator of the methods existed in deployment . The proposed measures are presented in Section 4 , and finally conclusions are content .

II. critical success factors of ERP systems

Implementing an ERP system is not an inexpensive or risk-free venture. In fact, 65% of executives believe that ERP systems have at least a moderate chance of hurting their businesses because of the potential for implementation problems .It is therefore worthwhile to examine the factors that, to a great extent, determine whether the implementation will be successful. Numerous authors have identified a variety of factors that can be considered to be critical to the success of an ERP implementation [3].

Based on empirical studies of ERP implementations in organizations for successful implementation of integrated ERP systems, project management, educating and culturing, setting project goals and scope, conforming with

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

organizational processes, software configuration, project team, development and defender of project methodologies and etc are important factors. The following are explained some of the factors successful.

A. Top Management Support/Commitment

Top management support was consistently identified as the most important and crucial success factor in ERP system implementation projects (Welti, 1999; Davenport, 1998a; Sumner, 1999; Bingi, et al., 1999; Gupta, 2000; Bancroft, et al., 1998).

Welti (1999) suggested that active top management is important to provide enough resources, fast decisions, and support the acceptance of the project throughout the company. Jarrar, et al. (2000) pointed out that the top management support and commitment does not end with initiation and facilitation, but must extend to the full implementation of an ERP system [4].

B. Define clear goals and objectives

ERP implementations require that key people throughout the organization create a clear, compelling vision of how the company should operate in order to satisfy customers, empower employees, and facilitate suppliers for the next three to five years. There must also be clear definitions of goals, expectations, and deliverables. Finally, the organization must carefully define why the ERP system is being implemented and what critical business needs the system will address [3].

C. Business Plan and Vision

A clear business plan and vision should be behind the implementation strategy to know in which direction the project must be steered. In project management three often competing and interrelated goals that need to be met are mentioned: scope, time, and cost goals. There must be a clear business plan how the goals can be achieved.

Business plan and vision summarises the CSFs clear goals and objectives, management (of) expectations, anticipated benefits from ERP implementation project, business plan and vision, adequate ERP implementation strategy, motivation behind ERP implementation, multi site issues and business case [5].

D. Project Team

ERP implementation teams should be composed of top-notch people who are chosen for their skills, past accomplishments, reputation, and flexibility. These people should be entrusted with critical decision making responsibility .Management should constantly communicate with the team, but should also enable empowered, rapid decision making .

The implementation team is important because it is responsible for creating the initial, detailed project plan or overall schedule for the entire project, assigning responsibilities for various activities and determining due dates. The team also makes sure that all necessary resources will beavailable as needed [3].

E. Communication

As the goal of ERP systems is to integrate various business functions across different locations, interdepartmental cooperation and communication is the core of the ERP implementation process (Akkermans and Helden, 2002), he suggested that intensive communication between the key parties is directly linked to the success of the project [6].

III. Methods of deployment ERP systems

One of the main points that exist in implementation of large information systems, utilizing an approach and methodology of development and implementation of system. The importance of this in ERP system is the nature of ERP systems and impacted all processes and activities of the organization.

Methodologies are crucial for the deployment of ERP systems can refereed AIM method that is created by Orale company, ASAP methodology by SAP company and Signature by Epicor Scala company. The large firm's producer using a specific methodology for the deployment of ERP systems. ASAP methodology has five phases that is a comprehensive and rich approach and, significantly reducing the overall cost and quality of the work is done at a high level [7]. In this method there are support from project management, member of team, external consultants and technical consultants, business process [7]. and a great tool for small and medium businesses [8]. Project management is a critical factor in the implementation of ERP systems, is provided by the ASAP. Good project management, especially in the process of designing, testing and end user training are important factors in successful implementation by SAP in the most organizations [6]. ASAP is a fast and flexible methods [7,8].

Methodology provided by the company Oracle is named Application Implementation Method (AIM). This methodology involves defining the activities, work processes, standards, procedures and practices, that detail of it described in six different sections with Milestone set guide and valuation activities relative to each other and define the main activities for the speedy implementation of projects and activities are complementary choice. In order to successfully implement this methodology, first action is required, and the resources needed to do and the resource needed for accomplish a specific project are recognized, and secondly, to do all of the activities, provides a patterns for the outputs . The main advantage of this methodology is that business requirements are defined early in the project and during implementation Consider placed. One of the major disadvantages of this methodology, its complexity [7]. Framework that are including elements such as steps, processes and tasks. AIM has a very wide scope, in this field investment of firms, sectors and there is a group of branches [9].

Signature methodology used in small and medium businesses [10]. Reducing the overall cost of the system and the attitude same all of Scala consultants in all areas project. From aspect of learning provides set of standard Classroom and Training Web-based.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Methodology ASAP

Figure 1 shows the phases in ASAP methodology

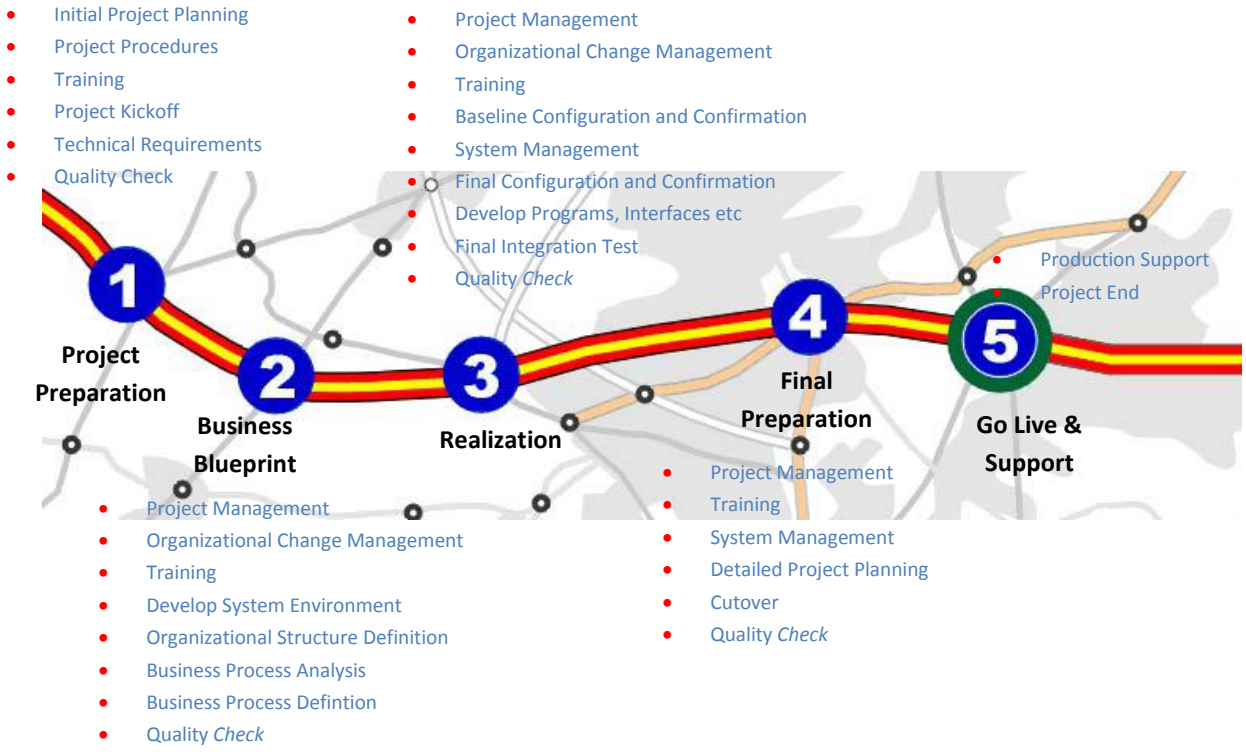


Figure 1: phases in ASAP methodology [12]

Phase I - Project Preparation

In this phase is done the preparation of elementary tasks of project [11]. Also in this step preparation of project charter, create Structure and organization of project review and refine the plan and implementing strategy implementation solutions, creating work teams and assign tasks, create plan and detailed action plan , defining technical requirements , initial meetings for bing project identification process model for doing , modeling and analysis of project requirements and determine organization of project , etc Is performed [12].

Phase II - Business BluePrint

Review, identify and design business processes in different fields is done via a standard procedure at this stage. In additional review and modeling of the business objectives and existing structures that are done [13]. In this phase after modeling current processes of organization, achieve to step modeling the future state of organization after changes [8].

Phase III – Realization

In this phase, tasks such Training project team, initial configuration of the system and receive confirmation, changes in basic ERP software has become an appropriate ERP solution to customize by means defined, creating and testing necessary interfaces, creating reporting Tools and testing them, testing integrity of the system is performed. Also in this phase program for transition is regulated [8,12].

Phase IV - Final Preparation

This phase allocated final preparation and review the plans projects. In this phase, system administration and user training, final testing of the system, applies the modifications and changes, transfer data from old systems to new systems is performed [8,12].

Phase V: Go Live & Support

Preparation and review launching System, correcting errors, preparing plans and schedules of timing and supporting its and activities of closing project in the final phase will be performed [8,12].

B. Methodology AIM

Figure 2 shows the phases and operations of the AIM methodology

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

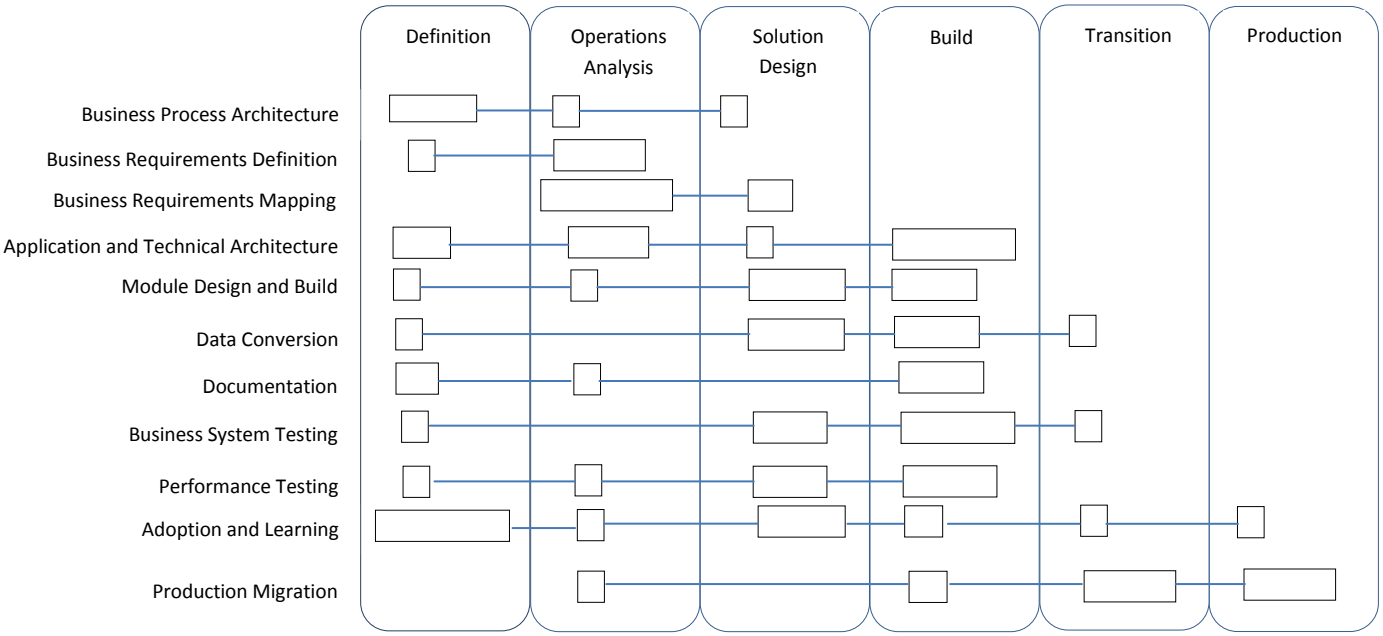


Figure 2: phases of the AIM methodology [9]

Phase I: Definition

In this phase, the employer and Contractor working together for planning process, review the resources and constraints, project scope and organization of operational teams [9,14].

Phase II: Operations Analysis

In this phase, the project team will move towards collecting work processes which can be extract differents current work step with standard applications ERP. Also, decide on the future organization of work processes are performed at this stage [9,14].

Phase III: Solution Design

The purpose of this phase detailed design new solution for the business needs and organization. Also based on organization's needs and if it is optional, it can be added selection other features can be added in solution. In this phase based on decisions of phase II provided solutions for conforming current work processes with standard process ERP [9,14].

Phase IV: Build

Coding and testing all areas of custom software, conforming software application in the organization, conversion data and design user interfaces is formed in this phase. effective testing and System testing is done in this phase. The purpose of this phase is to formulate and provide detailed requirements for computer applications and present a solution [9,14].

Phase V: Transition

At this stage programs are made in previous phase the implemented operationally in organization and the data in the system before being transferred to the new systems and its weaknesses, is amended as. In this phase, the current business processes and ERP applications are working parallel. In other words application programs are tested in a real environment [9,14].

Phase VI: Production

The final delivery of the new system at last phase of this methodology and the beginning of system support cycle has been done. Improvement and steps of measurement to be carried out at this stage [9,14].

C. Methodology Signature

Figure 3 shows the phases in the methodology Signature.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

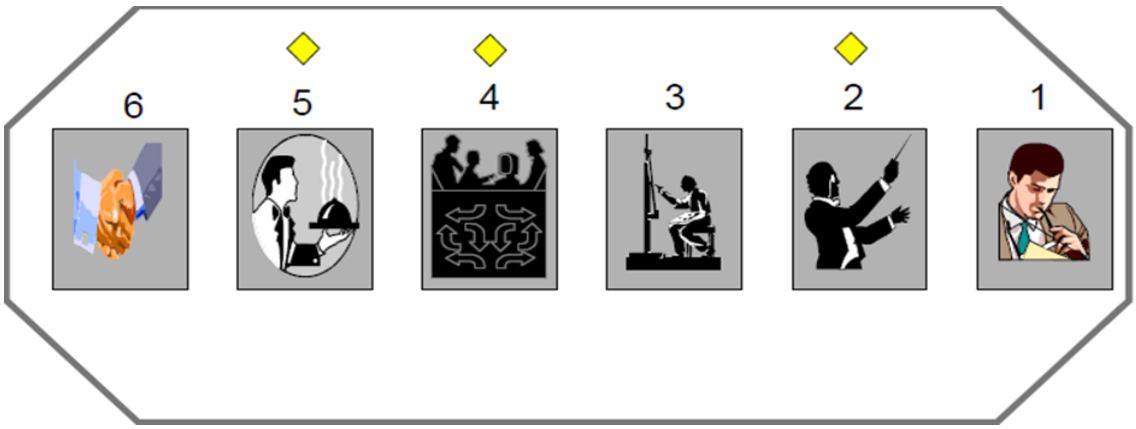


Figure 3: phases of methodology Signature.

Phase I: Analysis

The objectives of this phase definition of client's business requirements, identify key individuals of the project team, and the design of the prototype system. This phase will evaluate business requirements, key members of the project team trained [15].

Phase II: Project Organization

At this stage , the project objectives are set and approved, project plan preparation, proofing and approval, the project team and its schedule is determined. Project management and control procedures are in phase [15].

Phase III- Design

The purpose of this phase is to design and configure the system. Design business strategies, design inputs, outputs and intermediate users set the parameters of the test system and training end-users in all phases of the system are achieved [15].

Phase IV - Data Preparation

This phase of the procurement, transfer, and data validation has been allocated. Activities in this phase include: Define data conversion requirements, and construction and determine conversion methods, data convert to the new system and ensure the integrity of transmitted data [15].

Phase V - Test Run

The purpose of this phase is to review the project to ensure the proper functioning of the system . In addition, the pilot implementation of the system according to the requirements of business and the final configuration phase takes place in the system [15].

Phase VI – Hand Over

The final phase is allocated to operating new system, closure and delivery of projects. Operating system, project evaluation, quality control, delivery to support team will do well in this phase [15].

There are important indicators in each of methodologies presented. in the major indices are for the implementation of ERP systems can refereed to planned and project management, business process re-engineering, training and test and quality control.

Methodologies to deployment an integrated ERP system is an key success factor. And there are indicators in each of these methodologies that this is one of the success factors. For example, if the planning index is not, The project will most likely fail. Although the methods are deployed with the same parameters But the difference in one or more indicators is caused better success in the implementation of projects and minimizes the risk of project failure.

Project on Management is one of the most important indicators. This includes change and communication management and systems management and etc. Good project management in the process of designing, testing, and training, etc, one of the important factors is for successful ERP implementation in the organization. Table 1 lists the most important indicators of the ways it shows.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

	Index	AIM	ASAP	Signature
1	programming	✓	✓	✓
2	Fast implementation	✓	✓	
3	Customization	✓	✓	✓
4	Project management	✓	✓	✓
5	Business Goals	✓	✓	✓
6	Risk reduction	✓	✓	
7	Support for complex projects	✓		
8	Education	✓	✓	✓
9	Architecture SOA	✓	✓	✓
10	Test	✓	✓	✓
11	Reengineering	✓	✓	✓
12	Configuration	✓	✓	✓
13	Repeatability	✓	✓	
14	Preparation	✓	✓	✓
15	Business more efficient	✓	✓	✓
16	Flexibility	✓	✓	
17	Time reduction	✓	✓	
18	Cost reduction	✓	✓	✓
19	Complete coverage of the entire project life cycle		✓	
20	Designing TO-Be		✓	
21	Designing As-Is	✓		✓
22	Implement a comprehensive and powerful		✓	
23	Clear definition of project scope	✓	✓	
24	Incremental Model	✓	✓	
25	Support	✓	✓	✓
26	Help Desk		✓	
27	Documentation	✓	✓	✓
28	Multi-site	✓	✓	✓
29	Multi nationality - linguistic	✓		

Table 1: important indicators of ERP

IV. Proposed Method

In section 3 for the implementation of ERP systems indexes methodologies are studied and evaluated. In this regard, after reviewing the existing indicators, In the methodologies to deployment we recommended indicators for successful implementation of the systems ERP.

Security Index - Since ERP systems with resources and information on the organization and sometimes outside agencies interact in the matter and determine the boundaries of information security important to access this information. As a result, information security, and set limits for access to this information is important. Since the different layers vary in different organizations and the type and amount of information they are different. when implementing ERP systems in the organization should be a good strategy to control access to systems and data should be considered . Security software is a software quality assurance activity that focuses on the detection of potential risks and may have a negative impact on software as well as ruining the entire system, under three headings to integrate security, privacy and accessibility is divided. Due to that none of the methods proposed methodology is no way to control security. The proposed method, called security control measures in the will proposed methodology. That methodology will be presented in the strategy and planning and management security controls. Simultaneously with the planning and management of security controls to control costs.

Indicators of Quality Control - Quality control is a set of open projections, review and test the process used software product in order to ensure compliance with the requirements. For which it is defined. Quality control includes feedback loops, a process that has created a working product. Quality control parameters for planning, management, methods used for quality control and testing to detect errors in the different stages. Table 2 shows the parameters in the proposed method.

	Index	AIM	ASAP	Signature	proposed method
1	programming	✓	✓	✓	✓
2	Fast implementation	✓	✓	✓	✓
3	Customization	✓	✓	✓	✓
4	Project management	✓	✓	✓	✓
5	Business Goals	✓	✓	✓	✓
6	Risk reduction	✓			
7	Support for complex projects	✓	✓	✓	✓
8	Education	✓	✓	✓	✓
9	Architecture SOA	✓	✓	✓	✓
10	Test	✓	✓	✓	✓
11	Reengineering	✓	✓	✓	✓
12	Configuration	✓	✓	✓	✓
13	Repeatability	✓	✓		✓
14	Preparation	✓	✓	✓	✓
15	Business more efficient	✓	✓	✓	✓
16	Flexibility	✓	✓		✓
17	Time reduction	✓	✓		✓
18	Cost reduction	✓	✓	✓	✓
19	Complete coverage of the entire project life cycle		✓		✓
20	Designing TO-Be		✓		✓
21	Designing As-Is	✓		✓	
22	Implement a comprehensive and powerful		✓		✓
23	Clear definition of project scope	✓	✓		✓
24	Incremental Model	✓	✓		✓
25	Support	✓	✓	✓	✓
26	Help Desk		✓		
27	Documentation	✓	✓	✓	✓
28	Multi-site	✓	✓	✓	✓
29	Multi nationality - linguistic	✓			✓
30	Quality control				✓
31	Security				✓
32	Supervision				✓

Table 2. List of parameters in the proposed method

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. Conclusion

ERP is a software solution to integrate enterprise resources is beneficial. In this regard that the goal of software engineering to produce quality products in accordance with cost control is considered, therefore, to achieve this requires strategy and planning and management methodology that processes for quality control in should be considered. And implementation of a framework for the process to be used, Security software is a software quality assurance activities, in this regard that the methodologies is an important factor in the deployment, so if the security controls and control procedures should be implemented, increases the likelihood of successful implementation of ERP systems in organizations.

References

[1] A.Tambovcevs, and Y.Merkuryev, "ANALYSIS OF ERP SYSTEMS IMPLEMENTATION IN THE CONSTRUCTION ENTERPRISES," Information Technology and Management Science, 2009.

[2] D. Litan, and A. Apostu, and L. Copcea, and M. Teohari, "Technologies for Development of the Information Systems: from ERP to e-Government," INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS, vol. 5, pp.137, 2011.

[3] J. Umble, and R. Haft , and M. Michael Umble," Enterprise resource planning: Implementation procedures and critical success factors," European Journal of Operational Research 146 (2003)241–257, pp.244-246 , 2003.

[4] M.Al-Mashari, and M. Zairi ,and A.Al-Mudimigh, "ERP Implementation: An Integrative Methodology," RP—ECBPM/0012, pp.8, 2010.

[5] A. Kronbichler, and H.Ostermann, and R.Staudinger," A Review of Critical Success Factors for ERP-Projects," The Open Information Systems Journal, Vol.3, pp.16 , 2009.

[6] K.Elmeziane, and M.Elmeziane,"Enterprise Resources Planning Systems Implementation Success In China," Business and Management Review, Vol. 1(12), pp. 5, 2012.

[7] D.Manojlov, and MiroslavLutovac,"IMPACT OF METHODOLOGY IN THE SUCCESS OF ENTERPRISE RESOURCE PLANNING (ERP) IMPLEMENTATION," INFORMATION TECHNOLOGIES IN MANAGEMENT, pp.886-891 ,2012.

[8] J.Esteves, and Joan A. Pastor," A FRAMEWORK TO ANALYSE MOST CRITICAL WORK PACKAGES IN ERP IMPLEMENTATION PROJECTS," International Conference on Enterprise Information Systems, 2002.

[9] "APPLICATION IMPLEMENTATION METHOD HANDBOOK," August, 1999, Part Number A75149-01

[10] I.Osobskaya, and V.Skaredov," INTEGRATED INFORMATION SYSTEMS AS A FACTOR OF RELIABILITY OF LOGISTIC SYSTEMS IN SUPPLY CHAIN ORIENTED BUSINESS," 5th International Conference RelStat'05, Vol.7, No 1, pp.173 , 2006

[11] A.Momoh, and R.Roy, and E.Shehab, "A Work Breakdown Structure for Implementing and Costing an ERP Project," Communications of the IBIMA, Vol.6, pp.94-95 , 2008.

[12] "ASAP91 SAP Implementation ASAP91," R/3 System Release 46C 04/05/2001.

[13] A.Elragal, and M.El Kommos," In-House versus In-Cloud ERP Systems: A Comparative Study," Journal of Enterprise Resource Planning Studies, Vol. 2012, pp.5-7 , 2012.

[14] C.Liu, and L.S. Sally Chen," APPLICATIONS OF RFID TECHNOLOGY TO IMPROVE PRODUCTION EFFICIENCY FOR INTEGRATED-CIRCUIT ASSEMBLY HOUSES," 19th International Conference on Production Research.

[15] URL of EPICORE: www.epicor.com

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

RFID Technology: *Analytical Study using SWOT and STEEPLE Approach*

Rana Ibrahim Alabdan/ Lecturer
Information Systems Department. Majmaah University
Riyadh, Saudi Arabia

***Abstract*— The benefits of RFID technology cannot be denied and the new evolution in retails, supply chain management and companies when using this technology is really relevant and make the operation of production smoothie and easy. However, RFID also has some negative issues such as violation to the people's privacy and violation to the data protection. Even though, these security issues RFID has changed the way dealing with products and people. Thus, let's take this opportunity, improve the life, use new technology even if there is a little fair do not hesitate to take the benefit from every new technology that change our life to the best.**

***Keywords-component; RFID; IT technology; Networks; RFID issues; tags; SWOT Introduction* (Heading 1)**

I. INTRODUCTION

RFID stands for Radio Frequency Identification. The main goal of an RFID system is to carry data on a transponder (tag) that can be retrieved with a transceiver through a wireless connection. The ability to access information through a non-line-of-sight storage in a tag can be utilized for the identification of goods, locations, animals, and even people. Discerning specific information from these tags will have profound impacts on how individuals in commerce and industry keep track of their goods and each other. Early use of this technology concerned the evolution of barcode applications, changing the application scenario perspective [4]. The acronym RFID, Radio Frequency Identification, encompasses a number of technologies usable to identify objects by means of radio waves. The origin of the technique is the "Identification Friend or Foe" IFF system used in World War II by the Royal Air Force, that was able to get a code back only from "friendly" aircrafts identified with RADAR. Under this very wide umbrella the term is today mainly referring to systems where electronic equipment can "read" information from a multitude of "tags" by means of radio waves. The RFID tag can come in various shapes e.g. as a paper sticker, just as barcode tags are, as a plastic Credit Card, or even as a rugged, chemicals and heat resistant, plastic capsule. The tag might be even powered by a very small battery to support local functions such as storing temperature readings or enhancing the reach of the radio communication [5].

Although RFID is a mature technology, it took several years for a large-scale implementation to occur. The first ones were in the United States. The implementation eventually included supply chain, freeway tollbooths, parking areas, vehicle tracking, factory automation, and animal tagging. The most common application of RFID technology today is for tracking goods in the supply chain, tracking assets, and tracking parts from a manufacturing production line. Other application areas include the control of access to buildings, network security, and also payment systems that let customers pay for items without using cash.

Nevertheless some technology related issues still condition the possible applications. As an example, liquids, water especially, absorb radiations while metals reflect it. That means that passive tags applied to bottles of water or to aluminum cans can be hardly read though placed very carefully with respect to the reader antenna and with dielectric support. This is due to the properties of the radiations in relation to their wavelength It is true for HF tags but even more relevant for UHF tags [5].

The three basic components of a typical RFID system are an antenna or coil, a transceiver (reader with decoder), and a transponder (RFID tag) with electronically programmed information. In an RFID system, an antenna continuously emits radio signals at a given frequency. When a transponder comes into contact with these signals, the badge is activated and communicates wirelessly with the reader through the modulation of transmittance frequencies. Through the use of an antenna, the information that is stored on the transponder can be read or written from the transponder. Typically, the antenna is packaged with the transceiver into a larger structure called a reader that is in charge of the system's data communication and acquisition. The data that is obtained and analyzed by the reader can then be transported to a computer [5].

This paper consists of four major sections, which are:

1) Introduction
2) SWOT Analysis for RFID Technology
3) STEEPLE Analysis for RFID Technology
4) Conclusions

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

II. SWOT Analysis for RFID Technology

After RFID have been explained in details, definition, acronym and what is RFID, now it will be demonstrated from technology's perspective, what are the strengths, weaknesses, opportunities and threats. It has a great number of impacts toward technology in either way positive or negative that is applied for the companies supply chain and retailers and even so in the supermarket such as Wal-Mart or Pharmacy like Walgreen.

2.1 Strengths

RFID tags allow for easier logistics and product-related information storage and retrieval which could help a lot in making efficient, up-to- the-minute, event based management decisions. It also helps remove blind spots within the supply chain, which could help in eradicating some logistics inefficiencies such as product wastage/spillage, mishandling, etc.

In addition, a good tracking mechanism can help to ensure that the product being delivered within the supply chain are probably taken care of and are certainly transported to where they should be directed. It might help retailers ensure that they get what they really asked for instead of fraudulent products that could have been switched along the supply chain. It eliminates the inefficiencies brought about by theft, human error, and other fortuitous events that might hamper the over-all security of the goods.

RFID tags can help ensure a more efficient way of managing product inventories because of its inherent efficient ability in storing and retrieving product-related information. RFID Product information encoding is faster as compared to normal bar code systems and so is information retrieval because of its ability to store and retrieve information from multiple products at the same time. Overall, this will maximize inventory control and management [8].

2.2 Weaknesses

Although these tags can carry a lot of benefits, they are still very costly to implement because of the significantly high prices in buying the required amount of RPC RFID tags and receivers. RFID tags are useful in tracking products within the supply chain, but the receivers still have a proximity limit and thus each station must be cooperative in using the RFID technology [8].

2.3 Opportunities

RFID can give firms a lot of tracking and logistics related benefits, which makes it a very attractive investment for firms. Since these users usually buy these products in mass bundles this could be a very profitable product to focus on. The fact that there is currently a growing need for RFID technology in the modern world makes it a very profitable product.

Nowadays, the current prices for the cheapest RFID technology has significantly decreased by ten folds because of the current trending inverse relationship between demand volume and RFID cost. As the demand for more RFID technology grows the costs of installing RFID technology also becomes lower [8].

2.4 Threats

The entire supply chain process passes by a lot of different entities and the entire trip from growers to retailers can measure up to very long distances because of the far away placement of both the growing sector and retailing sector. This could make the logistics trip pass through a lot of entities and the seep in privacy brought about by RFID tags might not be welcome to these entities.

The current RFID tags could help improve overall efficiency of the companies in doing their work but the problem with this is that it could be a possible cause of laying off human resource, which are no longer needed because of the presence of new RFID technology. This could be bind smooth implementation of this technology [8].

III. STEEPLE Analyses for RFID Technology

3.1 Social

There are many major impacts from the company's perspective to RFID technology and employed it in the organization or in the market in general. Currently, a major drawback to widespread deployment of RFID systems is the overall attitude of people towards them. These days, social acceptance and trust of RFID is quite low as a result of insufficient privacy and security safeguards and also the lack of awareness and may retard take up of RFID technology.

3.2 Technological

To have RFID technology inside the company they should know the best way to setup this technology and thus to have the right goals from being installed. Sometime, high initial costs for setting-up RFID systems, uncertainty on the future for the RFID technology, the lack of well established standards and finally hidden societal, and organizational costs are well-known barriers for smaller companies which are reluctant to adopt this technology. Sometime, collision between RFID readers may happen when two signals are interfered [1].

3.3 Economical

It will create new jobs, related to data processing and service-related jobs as a result economic growth may also contribute to the creation of additional workplaces. It is clear that training activities will be needed as new kinds of skills will be required both for professional workers and for the end users which will affect the company's economic because they need to pay for all these kind of training and new jobs were created [9].

3.4 Environmental

RFID technology will play a critical role in helping companies determine the environmental impacts of manufacturing and recycling their products. RFID tags play a

Identify applicable sponsor/s here. *(sponsors)*

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

positive role in helping companies reduce the environmental impacts of their products' disposal. In other words, there is an opportunity for companies to use the visibility that RFID provides to look at the materials they use, their manufacturing operations, their supply chain and so forth, in order to reengineer products and processes to reduce costs and the impact their products have on the environment [5].

3.5 Political

Politicians should not be deciding how to design and implement RFID technology, from either efficiency or a security perspective. That role should go to those involved in creating and using this technology, and it's time for technology experts to work on that and start talking about it. If they do not, political agendas could get in the way of a positive future for RFID. According to Sonia in her article, "The senator from North Dakota is not alone in his interest. Lawmakers in at least five states are considering RFID legislation." They have fear from this new technology and they want to make a new regulation for RFID, which does not make any negative impact for anyone until now and the politician should give this technology a little opportunity [7].

3.6 Legal

RFID may have some possible legal issues in business such as contravention of the right to privacy and data protection, contravention of the right to personality, contravention of the right to human dignity, unfair competition and labor law violations. Thus, laws may be violated when RFID implements without taking civilian and business rights into consideration from the outset.

3.7 Ethical

As it is been known that the ethical issue is very significant one, there are some problems with RFID tags and readers. The content of an RFID tag can be read after the item leave the supply chain. Another issue is RFID tags are difficult to remove because they are very small and some of them are hidden inside the products. Moreover, RFID tags can be read without your knowledge. The most serious issue is RFID tags that have unique serial numbers could be linked to an individual credit card number [6].

IIII. CONCLUSIONS

In conclusion, the benefits of RFID technology can not be denied and the new evolution in retails, supply chain management and companies when using this technology is really relevant and make the operation of production smoothie and easy. However, RFID also has some negative issues such as violation to the people's privacy and violation to the data protection. Even though, these security issues RFID has changed the way dealing with products and people. Thus, lets take this opportunity, improve the life, use new technology even if there is a little fair do not hesitate to take the benefit from every new technology that change our life to the best.

References

[1] Chu, T. (2010, July 9). 3 Future Improvements of RFIT Technology. Modern Safety. Retrieved April 7, 2011, http://blog.fieldid.com/2010/07/3-future-improvements-of-rfid-technology/

[2] Murata to exhibit innovative RFID solutions for electronics products at CeBIT - Retail Technology Review. (2011, Feb 14). The Retail Technology Review - Retail supply chain and in store technology news. Retrieved April 7, 2011 http://www.retailtechnologyreview.com/absolutenm/templates/retail_rfid.aspx?articleid=1525&zoneid=2

[3] Softpeia. (2008, October 6). Retrieved from http://news.softpedia.com/news/California-RFID-Anti-skimming-Bill-Signed-into-Law-94992.shtml

[4] Tanenbaum, A. S., & Wetherall, D. J. (2012). Computer networks. (5 ed.). Boston: Prentice Hall.

[5] Lieshout, M., Grossi, L., & Spinelli, G. (2007). Rfid technologies: Emerging issues, challenges and policy options. Informally published manuscript, European Commission.

[6] Technovelgy.com. (n.d.). Retrieved from http://www.technovelgy.com/ct/Technology-Article.asp?ArtNum=20

[7] Arrison, S. (2004, November 1). The politics of rfid. Retrieved from http://news.heartland.org/newspaper-article/2004/11/01/politics-rfid

[8] Mikkaela, Y. (2012). Scribd. Retrieved from www.scribd.com/doc/61865049/Case-Exam-Savi-RFID-Swot-Analysis

[9] Study on the competitiveness of the eu security industry. (2009).

AUTHORS PROFILE

Author – Rana Ibrahim Alabdan, Master of Information Systems Management from Robert Morris University, Moon Township, PA 2012. Lecturer in Majmaah University, Information Systems Department, Majmaah, Saudi Arabia.Bachelor of Science in Computer Information Systems from Imam Mohammed Ibn Saud Islamic University 2008. CIS Outstanding Graduate Student Award –Robert Morris University (April 2013). Member in Alpha Iota Mu since (5 May 2013).

Two Phase K-Nearest Neighbors Appoach

Siddhartha Kumar Arjaria[1], Deepak Singh Tomar[2], Devshri Roy[3]
Department of computer science & Engg.
Maulana Azad National Institute of Technology
Bhopal, (M.P.), India
[1]arjarias@gmail.com, [2]deepaktomar@manit.ac.in, [3]droy.iit@gmail.com

Abstract— **K-nearest neighbors approach is the popular algorithm for classification. The majority of votes of neighbors of testing sample decide the class of in K-nearest neighbors approach. It only utilizes the information stored in the first few samples while it considers the remaining samples unimportant. The classification result of K-nearest neighbors approach highly depends on the single criteria, due to this classifier many times produces the wrong result. The paper presents a novel idea to deal with the classification problem in two Phases. First phase deals with the extraction of useful information from the training space regarding the occurrence behavior of each training sample in the neighbor list of other training samples. This occurring behavior decides each training sample to be part of one of the three classes namely important, unimportant, and neutral. In the second phase, On the basis of this collected information the training samples in the neighbors of testing sample are rearranged by removing the unimportant samples. Now classification decision totally omitted the unimportant training samples and considers only the important & neutral class training samples. Algorithm is designed to provide the extra weights to the important samples on the basis of its position in neighbor list, it's occurrence frequency as a neighbors of other training samples and the number of training samples of that class used for training. Performance is tested on three database seven most frequent categories of Reuters-21578, four most frequent categories of RCV1, seven most frequent categories of TDT2 corpus. our approach outperforms K-nearest neighbors approach in terms of F1 value in almost each case.**

Keywords- K-nearest neighbors approach*;*Two Phase KNN;Classification;

I. INTRODUCTION

In the present era of information technology, digital text information is growing exponentially. As a consequence, one of the major challenges is to organize and manage them automatically. One of the ways to organize text information is the classification of text documents into the correct category. Machine learning approaches are used for text classification. Different machine learning algorithms such as inductive learning [2], neural networks (W., E., J.O. Pedersen, and A.S. Weigend 1993) Naïve Bayes classifiers [14] decision trees, K-NN[7], Support Vector Machine (SVM) is already proposed .
The traditional KNN has some limitations. All the training samples are treated equally and given the same weight. But in practical situations, some samples are more significant than others. The contribution of more significant samples should be more as compared to others. The most significant sample could help in correct classification, which in turn increases the classification performance.
The approach used in this paper is divided in to two phases. First phase collects the information regarding the usefulness of training samples. For this the occurring behavior of each training sample in the neighborhood of other training samples give important information by finding how much time the Tri will be in the neighbors of training samples of class s. On the basis of the information each training sample will belong one of three classes namely important, unimportant and neutral. The important class training samples have the property that the occurrence of these samples is restricted to specific class. And their occurrence count in that class is above predefined threshold .They will almost not occur in the neighbor list of training samples of other class. The unimportant class holds the training samples that are occurred in more than one class with significant occurrence count. The neutral class has the remaining training samples which are not the part of important and unimportant class.
In the second phase for classification of each testing sample the information collected from phase1 is used. The neighbor list of testing samples contains the training samples that belong to one of three classes. It is clear that the occurrence of training samples of the unimportant class in the neighbor of testing sample have a confusing behavior of class occurrence (occurred in more than one class)so it's occurrence in the neighbor of testing sample need to be eliminated for good classification results. On the other hand the occurrence of neighbors of important class needs the encouragement as it's occurrence is dedicated to a single class only. The idea is to discourage the occurrence of training samples of unimportant class in neighbor list of testing sample by completely removing it from the neighbor list & in turn not consider it in class decision making. In place considering the next important or neutral neighbors of that testing sample, means that class decision making is completely based on samples of important and neutral class. For best classification result it is needed to increase the weight of important neighbors by a certain calculated factor. The weighting factor of each important samples determined by considering the following points

- The important neighbors that are near to testing sample provide more weights as compare to important neighbors that are far from the testing sample.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- The important neighbors that have more occurring frequency in training space provide more weights as compare to important neighbors have less occurring frequency in training space.
- If the important neighbors belongs to the class that have more samples for training will provide the less weights as compare the important neighbors that belongs to the class that have less samples for training

Combination of all these factor leads towards the increased classifier performance as compared to KNN classifier algorithm.

II. Related Work

[8] had given K-nearest neighbor approach for text classification Traditional KNN is one of most popular and extensively used method of text classification. However it has many drawbacks like classification accuracy depends on the value of K, equal value of weight to K nearest neighbours, no differentiation of relevant features, the classification performance is easily affected by irrelevant features, more memory requirement etc. To overcome the drawbacks of KNN, many researchers had worked & proposed some improvement.

A weighted K-nearest approach to overcome limitation of assigning equal weight to k nearest neighbors[3] is proposed. Different Weights are assigned to all training samples according to the distance calculated. All training samples are used instead of K nearest neighbors. Therefore, computation complexity is increased in calculating weights for all training samples made the algorithm slow. [1] introduces Condensed nearest neighbor. It Reduce size of training data , by elimination of training data which do not add new useful information due to this query time and memory requirements improves. [15] proposes a method nearest feature Line Neighbor uses the ignored information of nearest neighbor and improves the accuracy this is good for small size but the computational complexity of this approach is increased. [19] propose a modification in KNN which uses the concept of weights and validity of data point for classification. The low accuracy of weighted KNN is improved by them partially. [9] proposed reduced nearest neighbor method for text classification. The aim of this method is to reduce the size of the training data by eliminating patterns which do not affect the final result. Advantage of this method is less memory requirement. The disadvantages are computational complexity; cost and time consumption is more.

[10] proposed an improved KNN algorithm for text categorization, which builds the classification model by combining constrained one pass clustering algorithm and KNN text categorization. [23] proposed an improved KNN text classification algorithm based on clustering center. It clusters each training set by K value clustering and obtain all cluster centers from the new training set. According to the number of training samples in each cluster, weight is assigned. The training samples which lie near border are removed from the training sets. However, these samples also give important information that the border training samples have dominance of the other classes than the actual class. This fact can be used in the decision rule for predicting the class of unknown training sample. Approach use in this paper achieves the significant improvement in accuracy as compared to the traditional KNN.

III. Two Phase KNN

The algorithm is break in two phases. the output of phase1 will work as the input of phase 2.

A. Phase1

The phase 1 starts with the feature selection of each document

1) Feature selection

The steps used for feature selection is discussed in algorithm1

Algorithm 1. Feature selection

Input

1. Document set of n documents
2. Stop word list
3. Value of local and global threshold
4. N: Number of features selected

Output

1. Term Frequency matrix for $Document_i$

1. Begin
2. For i = 1 to Number of documents
3. Extract the term set satisfying local and global threshold
4. Remove the stop words
5. Apply stemming
6. Apply Normalization
7. Term Frequency matrix for $Document_i$
8. Next i
9. End

Up to this point the algorithm makes the term frequency matrix of each document. Although size of raw document is reduced by great extent, but all these terms are not equally important for classification. Terms should be picked up for in such a way that most specific and unique terms of different classes takes the leading edge over other terms, to boost the classifier performance in terms of accuracy, classification time and memory requirement.

In the last few years the researcher makes lot of attempt for choosing the best terms of document database. Few are odds ratio[17], document frequency[22], mutual information[16], information gain[11], improved Gini index[21], measure of deviation from Poisson distribution[18], chi-square[4]. The algorithm chooses the top N features using the measure of deviation from Poisson distribution to select the features for classification.

a) Degree of deviation from Poisson distribution

The measure of deviation from Poisson distribution method is used to select the features of databases. Hiroshi Ogura[18] proposed a feature selection method for text . It will take the degree of deviation from Poisson distribution in to consideration. The Poisson distribution is used for describing the probability of a number of events occurring in each series of units and it is given by

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$P(k\lambda_i) = \exp(-\lambda_i)\frac{\lambda_i^k}{k!}$$

Where λ_i is a positive real number which is equal to the expected number of occurrence.

$$\lambda_i = \frac{F_i}{N}$$

Where F_i is the total frequency of terms in all documents & N is the total number of documents in the text corpus.

$$POI = \frac{(A-\hat{A})^2}{\hat{A}} + \frac{(B-\hat{B})^2}{\hat{B}} + \frac{(C-\hat{C})^2}{\hat{C}} + \frac{(D-\hat{D})^2}{\hat{D}}$$

Where

A= Number of documents containing term t in documents of category c

B= Number of documents not containing term t in documents of category c

C= Number of documents containing term t and not belonging to category c

D= Number of documents not containing term t and not belonging to category c

$\hat{A}, \hat{B}, \hat{C}, \hat{D}$= predicted values of A,B,C,D

POI= Chi square statistic measure deviations from Poisson

M: Number of documents that belongs to category c

$\hat{A} = M.\{1 - \exp(-\lambda_i)\}$

$\hat{B} = (N - M).\{1 - \exp(-\lambda_i)\}$

$\hat{C} = M.\exp(-\lambda_i)$

$\hat{D} = (N-M).\exp(-\lambda_i)$

To collect information regarding the training samples, find the distances between each pair of training samples

Training space = Tr_i $1 \le i \le m$
Testing space = Te_j $1 \le j \le n$
Find the Euclidian distance between each training sample Tr_i to remaining training samples & arrange these neighbors for Tr_i in the ascending order of distance. Now select K1 neighbors of each training sample, thus finds neighbor list for all training set. Training space is distributed over c different classes. for the training samples of of class c ,find occurrence frequency of each training sample Tr_i in the K1 neighbors . Now for each training sample one of three cases will occurred

- The training sample Tr_i is dedicatedly occurred only in the class s of the c classes with frequency f. if f>threshold1 then Tr_i is in the imp_element of class s.
- The training sample Tr_i is occurred in the neighbor list of more than one class with the frequency f. if f>threshold2 then Tr_i is in the unimp_element maintained for all classes
- All other training samples Tr_i that are neither the member of imp_element of class s & nor the member of unimp_element is Part of neutal_element. All the Tr_i that are member of neutral_element either has no occurrence in any class c or have frequency below threshold.

Each training sample is the member of one of three above mentioned cases. This will improves the quality of neighbor list for each testing sample to be maintained in phase2.
NTr_i : m*K1 size array to hold the Neighborhood of Tr_i in training space
$FTTr_i$: 1*m size array to hold the Frequency of Tr_i in neighborhood of all training samples

Algorithm 2.Find the Important, Unimportant, Neutral element list
Input

- Training space: Tr_i $1 \le i \le m$
- represented by top N informative terms selected by deviation from Poisson distribution method
- K1:Number of neighbors $4 \le K1 \le 25$
- Th: Threshold value
- c: Number of classes

Output

- $imp\_element_s$: Array to hold the index of unique & most frequent training samples of class s
- $neutral\_element_s$: Array to hold the index of neutral training samples of class s
- unimp_element: Array to hold index of unimportant training samples
- $FCTr_{i,s}$: Array to hold the Frequency of Training sample Tr_i in neighborhood of training samples of class s

1. Begin
2. For i =1 to m
3. For j= 1 to m
4. Calculate Euclidian distance between i^{th} & j^{th} training sample
5. Next j
6. Sort distances in ascending order
7. NTr ← First K1 neighbors of Tr_i $4 \le K1 \le 25$
8. Next i
9. For s= 1 to c
10. For i= 1 to m
11. $FCTr_{i,s} \leftarrow 0$
12. for x=1 to m
13. If ($Tr_i \in NTr(x)$)
14. $FCTr_{i,s} = FCTr_{i,s} + 1$
15. Else
16. $FCTr_{i,s} = FCTr_{i,s}$
17. Next x
18. Next i
19. Next s
20. for i=1 to m
21. $FTTr_i = \sum_{s=1}^{c} FCTr_{i,s}$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

flag=false
For s= 1 to c
If$\left(\left(FCTr_{i,s} \geq Th\right) \;\&\&\; \left(\frac{FCTr_{i,s}}{FTTr_i} = 1\right)\right)$
$Imp\_element_s \leftarrow i$
flag = true
Break
Elseif$\left(\left(FCTr_{i,s} < Th\right) \;\&\&\; \left(\frac{FCTr_{i,s}}{FTTr_i} = 1\right)\right)$
$Neutral\_element_s \leftarrow i$
Flag=true
Break
Next s
if(flag = false)
unimp_element← i
Next i
End

B. *Phase2*

Work of this phase is to classify each testing samples Tej($4\leq j\leq n$)in class s($1\leq s \leq c$) The phase starts with distance computation of each testing sample Te_j with each training samples Tr_i. By arranging the distances in ascending order and build the neighbor list NTe_j for each testing Te_j. The neighbor list NTe_j contains the training samples tagged as one of the important, unimportant, and neutral in phase 1.

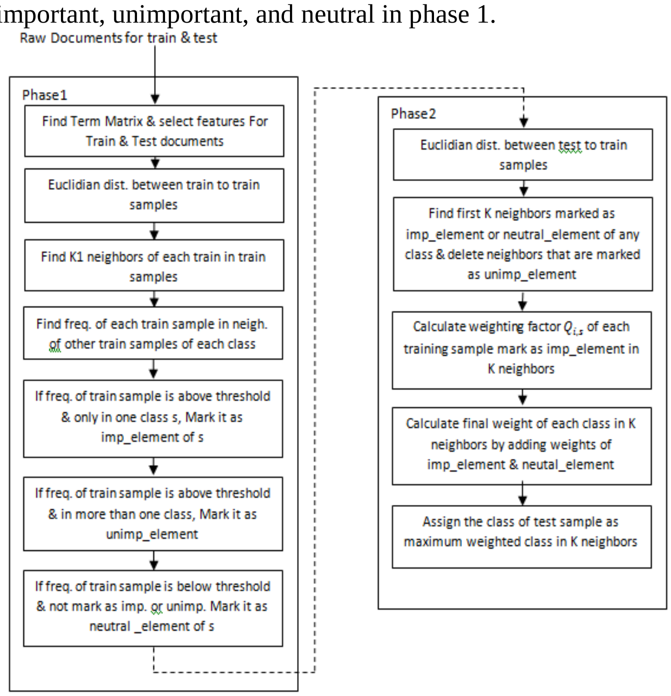


Figure 1. Two Phase KNN

Select of first K neighbors of important or neutral type in NTe_j by removing every unimp_element neighbors in between. unimp_element is removed because their behavior in phase 1 indicates that they have occurrence frequency in more than one class with above threshold value. So they are of confusing behavior in terms of class coverage. In such a way select K neighbors of important or neutral type for each testing sample.

Now if in neighbor list NTe_j , the neighbor Tr_i is the part of one of the imp_element of class s. From the phase 1 it is clear that the training sample in imp_element of class s is dedicatedly belongs to class s. So in classification process they strengthen the chances of occurrence of actual class. so, improve its weight by the factor Q .The weighting factor is designed in such way for neighbor of important type, so that it will leads towards the correct classification results. In the remaining neighbors, find the neutral_element type of class s & treat them in a non weighted fashion. In next step final weight for samples of each class appears in neighbor list NTe_j. is assigned to make the class decision.

Let out of K nearest samples the N samples belongs to class s. Out of these N samples Nimp (Nimp <=N) belongs to the imp_element category of class s and remaining Nneutral (Nneutral= N-Nimp) remains in neutral category of class s.

So the weight of Tr_i of class s in decision making:

$$W_s = \left(\sum_{i=1}^{i=Nimp} Q_{i,s}\right) + Nneutral_s$$

$$Q_{i,s} = \left(\frac{\max\limits_{1\leq i\leq m}(FCTr_{i,s})}{\sum_{i=1}^{i=m} FCTr_{i,s}}\right) * FCTr_{i,s} * \left(\frac{m}{m_{tr,s}}\right) * \left(\frac{K}{K_{mean}}\right)$$

$Q_{i,s}$:Weighting factor for important training sample Tr_i of class s in neighborhood of testing sample

$FCTR_{i,s}$: Frequency of Training sample Tr_i in neighborhood of training samples of class s

m:Total size of training space

$m_{tr,s}$:Positions of important training samples belonging to class s

K: Number of neighbors in decision making

$K_{tr,s}$:positions of important training samples among the K neighbors

K_{mean} :mean ($K_{tr,s}$)

The $Q_{i,s}$ or training sample i in class s depends on the frequency of occurrence in the neighbor list NTr_i. The ratio of maximum frequency of class s to total frequency of s class will give the value that maximum frequency has in total frequency. Now this will remain constant for the important training sample of class s. Now more the occurrence frequency of important training sample in NTr_i., more weight contribution of that training sample results.

The $Q_{i,s}$ for training sample is calculated in such a way so that it increases the weight of imp_element of class s in decision making. it also deals with the problem where there is huge difference in the number of training samples of each class s. Imagine the case where in total s classes the ratio of one class samples s1 to another class s2 is greater than 5 times then obviously the training samples of class s1 dominates in the neighbor list. In calculation of Q, the ratio between the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

total training samples to number of training sample belonging to class s will encourage the class s whose training samples are low as compared to number of training samples of other classes.

$Q_{i,s}$ for training sample i in class s also depends on its positions in the neighbor list of Te_j. The closer positions acquired by the important training samples of class s in the K neighbor of testing sample Te_j results in the higher weight participation in over all weight as compares to the important training samples that acquired the far positions. The $Q_{i,s}$ designed so encourages the possibility of correct class assignment to testing sample Te_j.

$$\text{class of testing sample: } c_{te_j} \leftarrow \arg\max_{1\le s\le c}\{W_s\}$$

Algorithm 3.Find class of testing sample

Input

1. Testing space: Te_j represented by top N informative terms selected by deviation from Poisson distribution method
2. $Imp\_element_s$:holds the index of unique & most frequent training samples of class s
3. $neutral\_element_s$:holds the index of neutral training samples of class s
4. unimp_element: holds index of unimportant training samples
5. $FCTr_{i,s}$:Frequency of Training sample Tr_i in neighborhood of training samples of class s
6. m: Total size of training space
7. $m_{tr,s}$:size of training samples belonging to class s
8. K: Total neighbors consider in decision making

Output

1. class of testing sample: c_{te_j}

Function Used

1. memeber(samples, Neighborlist) : Function that returns the index of element of matrix sample in the Neighborlist if they exist in it.

Variable Used

1. NTe_j : Neighbors of Te_j in training space
2. $NimpneuTe_j$: Contains K neighbors after removing the unimp_element for Te_j ($3 \le K \le 30$)

1. Begin
2. For j=1 to n
3. For i= 1 to m
4. Calculate Euclidian distance between i^{th} training sample & j^{th} testing sample
5. Next i
6. Sort distances in ascending order
7. selected_element = 0
8. $NTe_j \leftarrow$neighbors of Te_j
9. remain_element = K
10. do
11. I_{unimp}=search(unimp_element, NTe_j(remain_element))
12. selected_element=selected_element+(K-size(I_{unimp}))
13. remain_element=size(I_{unimp})
14. while((selected_element-K)~=0)
15. Put selected K elements in $NimpneuTe_j$
16. For s= 1 to c
17. I_{neu}=member($neutral\_element_s$, $NimpneuTe_j$)
18. I_{imp}=member($imp\_element_s$, $NimpneuTe_j$)
19. For r = 1 to size(I_{imp})
20. Ind= $I_{imp}(r)$
21. calculate $Q_{i,s} = \left(\frac{\max_{1\le i\le m}(FCTr_{i,s})*FCTr_{i,s}}{\sum_{i=1}^{i=m} FCTr_{i,s}}\right) * \left(\frac{m}{m_{tr,s}}\right) * \left(\frac{K}{K_{mean}}\right)$
22. Next r
23. $W_s = \left(\sum_{i=1}^{i=size(I_{imp})} Q_{i,s}\right) + size(I_{neu})$
24. next s
25. class of testing sample: $c_{te_j} \leftarrow \arg\max_{1\le s\le c}\{W_s\}$
26. Next j
27. End

IV. EVALUATION

F1 measure is used to evaluate the performance of algorithm. The F1 measure [20] is calculated on the basis of precision & recall measures

TABLE I. CONTINGENCY TABLE

	Relevant document	**Non relevant document**
Retrieved document	true positives (tp)	false positives (fp)
Not retrieved document	false negatives (fn)	true negatives (tn)

$$\text{Precision} = \frac{tp}{tp + fp} \quad \& \quad \text{Recall} = \frac{tp}{tp + fn}$$

$$F1 = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}}$$

The environment settings of computer is as follows: Intel(R)Core (TM) i3 CPU M 380 @ 2.53 GHz, 2.53 GHz, with RAM of 3GB, Operation System is Microsoft Windows 7.

A. Datasets

Performance of algorithm is evaluated on three bench mark data sets namely Reuters-21578, RCV1, TDT2 corpus.Reuters-21578 [5,13] collection has 21578 documents divided in to 135 categories. Seven most frequent categories in ‘‘ModApte” split of Reuter dataset is used. Reuters collections is widely used dataset in text categorization community [6].

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Reuters Corpus Volume I [12] is an archive of over 800,000 manually categorized newswire stories recently made available by Reuters, Ltd. for research purposes. The subset of RCV1 with 4 categories is used.

TDT2 [5]corpus consists of data collected during the first half of 1998 and taken from 6 sources, including 2 newswires (APW, NYT), 2 radio programs (VOA, PRI) and 2 television programs (CNN, ABC). It consists of 11201 on-topic documents which are classified into 96 semantic categories. In this subset, those documents appearing in two or more categories were removed, and only the largest 30 categories were kept, thus leaving us with 9,394 documents in total.

B. *Effect of variation of K1 (Neighbors in Train to Train) on average correctly classify samples*

The Fig. 2, 3 & 4 shows the effect of Value of K1 on the performance of our classifier in terms of correctly classified training samples of seven most frequent categories of Reuters-21578, four most frequent categories of RCV1, seven most frequent categories of TDT2 corpus. First fix the value of K1 and then for that value of K1 we perform the classification on the different values of K. All the K results we get in such a way is averaged to give the average training samples correctly classified for the particular K1 value. This is performed over all values of K1.

The behavior of graphs indicated that the value of K1 should not be too high or not too low. The too low values omit the some useful training samples while the large values of K1 include the unnecessary samples and chances to drop the classification accuracy.

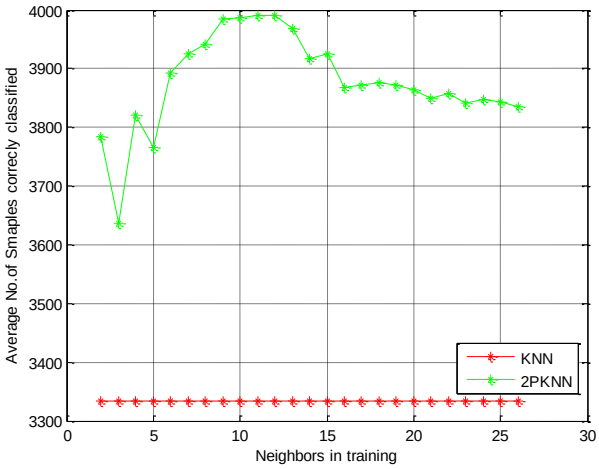


Figure 2. Classifier Performance analysis for different values Neighbors in training (K1) in terms of average sample correctly classified for RCV1

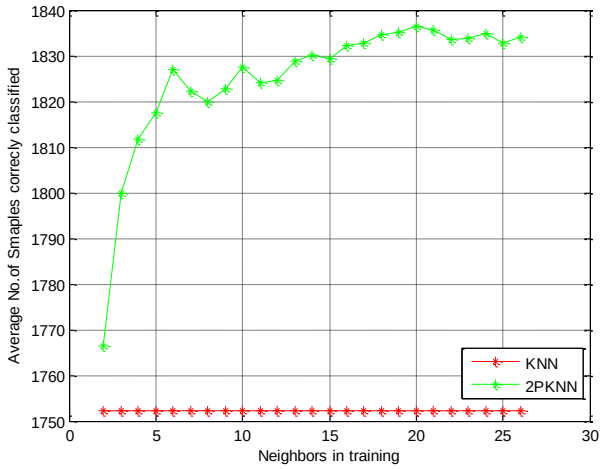


Figure 3. Classifier Performance analysis for different values Neighbors in training (K1) in terms of average sample correctly classified for Reuters-21578

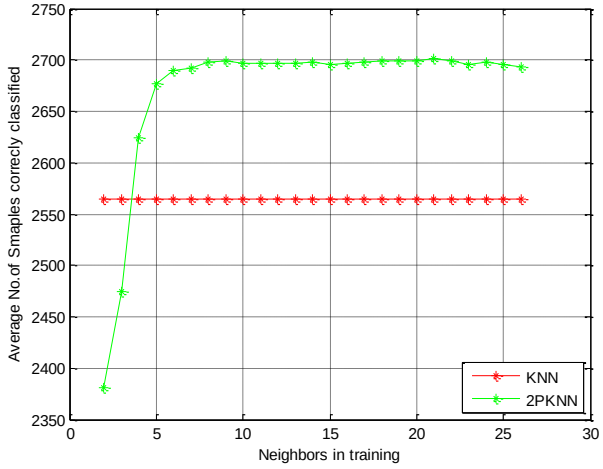


Figure 4. Classifier Performance analysis for different values Neighbors in training (K1) in terms of average sample correctly classified for TDT2

C. *Comparison of Two Phases KNN & KNN on Average F1*

The Table 2, 3 & 4 shows KNN classifier & Two Phases KNN performance for each category of database and also shows the detail of the number of training & testing samples for each category of Reuters-21578, four most frequent categories of RCV1, seven most frequent categories of TDT2. The features are selected by the method of deviation from Poisson distribution. The top 5000, 5500, 1200 terms having highest

TABLE II. COMPARING AVERAGE F1 VALUE OF KNN & TWO PHASE KNN ON DIFFERENT K1 VALUES FOR FOUR MOST FREQUENT CATEGORIES OF RCV1

Category	Train	Test	KNN	Two Phase KNN (F1 in %)					
			(F1 in %)	*K1=2*	*K1=5*	*K1=10*	*K1=15*	*K1=20*	*K1=26*
C15	1000	1022	50.38	68.17	66.94	60.90	60.16	59.56	58.99
ECAT	1000	1064	62.65	62.37	70.38	72.13	71.09	71.60	71.83
GCAT	1000	1901	57.35	71.10	63.98	73.73	73.23	71.10	70.31
MCAT	1000	1638	71.78	64.23	68.74	77.33	75.80	74.62	74.22

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE III. COMPARING AVERAGE F1 VALUE OF KNN & TWO PHASE KNN ON DIFFERENT K1 VALUES FOR SEVEN MOST FREQUENT CATEGORIES OF REUTERS-21578

Category	Train	Test	KNN (F1 in %)	Two Phase KNN (F1 in %)					
				K1=2	*K1=5*	*K1=10*	*K1=15*	*K1=20*	*K1=26*
Earn	2700	1013	92.60	96.72	96.54	96.39	96.25	95.96	95.80
ACQ	1400	655	88.84	90.11	90.07	90.89	91.17	91.88	91.61
Crude	221	100	67.64	82.23	81.01	83.12	84.39	85.21	85.54
Trade	200	98	64.95	79.72	79.75	79.59	79.49	80.09	80.27
Money-fx	180	65	48.52	51.40	50.12	55.20	52.62	56.94	57.71
Interest	150	47	70.16	67.27	68.94	67.77	69.06	73.92	74.83
Ship	110	32	39	37.03	36.37	46.82	45.99	45.81	43.48

TABLE IV. COMPARING AVERAGE F1 VALUE OF KNN & TWO PHASE KNN ON DIFFERENT K1 VALUES FOR SEVEN MOST FREQUENT CATEGORIES OF TDT2

Category	Train	Test	KNN (F1 in %)	Two Phase KNN(F1 in %)					
				K1=2	*K1=5*	*K1=10*	*K1=15*	*K1=20*	*K1=26*
20001	1100	744	89.73	81.10	93.97	95.34	95.50	95.63	95.44
20015	1100	728	93.71	93.30	98.25	98.34	98.00	98.04	97.91
20002	700	522	96.88	95.04	98.58	98.06	98.27	98.33	98.24
20013	500	311	88.83	84.04	96.44	96.66	96.67	96.67	96.49
20070	241	200	96.78	93.87	96.18	98.75	98.76	98.80	98.71
20044	240	167	98.06	68.67	98.44	98.63	98.50	98.42	98.54
20076	150	122	68.14	58.32	77.88	81.38	80.99	81.95	79.64

value of deviation from Poisson distribution are selected as a feature for Reuters-21578, RCV1, TDT2. The results shown in the tables are the comparison of F1 value for each class by Two Phase KNN & KNN. For a certain value of K1, find the F1 for each value of K (1to 18).and then averaged these values and get the averaged F1 value for that K1. In such a way averaged F1 value for all K is achieved. The table 2,3 & 4 indicates the classifier performance on K1=2,5,10,15,20 & 26. The table also shows the number of training & testing samples used in classification.

V. CONCLUSION

Every classification algorithm is designed keeping the accuracy enhancement in mind. This paper proposes certain well designed modifications in KNN that leads towards the better classification performance of classifier. The algorithm used in this paper extract the information about important, neutral & unimportant elements of each class from training samples. This information is extracted in phase1. The main factor involved in this phase is the number of neighbors for training sample, e.g.K1.The K1 should be neither too low nor too high.

In the second phase the class decision making depends on the filtered list of neighbors, means no unimportant training sample should present as the neighbors of testing sample. The K important or neutral training samples only are used as neighbors of testing sample. Each important training sample in these K neighbors is multiplied with a weight. The weight for each important training sample is calculated on the basis of

(1) Occurrence frequency in training space
(2) Its positions of important training samples among the K neighbors.
(3) The number of training samples in that important training sample's class

The neutral elements are processed in a conventional fashion by no increase in weight for class decision making. The performance is tested on three publicly available bench mark databases namely Reuters-21578, RCV1, & TDT2 corpus. In each case Two Phase KNN outperforms the traditional KNN approach in terms of F1 measure. This improvement in performance require more time for algorithm to find and use the additional information as compare to traditional KNN.

REFERENCES

[1] Angiulli ,F.: Fast condensed nearest neighbor rule, In: Proceedings of the 22 nd International Conference on Machine Learning (2005)

[2] Apte, C. , Damerau,F. ,Weiss, S. M .:Automate learning of decision rules for text categorization. ACM Transactions on Information Systems. 12(3), 233-251(1994)

[3] Bailey ,T., Jain, A. K.:A note on Distance weighted k-nearest neighbor rules.IEEE Trans. Systems, Man Cybernatics. 8, 311- 313 (1978)

[4] Chen ,Y.T., Chen,M.C. : Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications. 38 (4),3085–3090(2011)

[5] Dataset collection from http://www.cad.zju.edu.cn/home/dengcai/Data /TextData.html accessed on may 2013

[6] Debole, F., Sebastiani, F. :An analysis of the relative hardness of reuters-21,578 subsets. Journal of the American Society for Information Science and Technology.56(6), 584–596(2005)

[7] Duda, R.O. , Hart, P.E.:Pattern Classification and Scene Analysis. New York, John Wiley & Sons(1973)

[8] Fix, E., Hodges, J. L.JR: Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. USAF School of Aviation Medicine. Report No. 4, Project No. 21-49-004(1951)

[9] Gates ,G.W. :Reduced Nearest Neighbor Rule, IEEE Trans Information Theory.18(3),431-433(1972)

[10] Jiang,S., Pang, G., Wu ,M., Kuang,L.: An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications. 39,1503–1509(2012)

[11] Lee, C., Lee , G.G.: Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing and Management. 42 (1), 155–165(2006)

[12] Lewis ,D. D., Yang ,Y., Rose,T.G., and Li, F.:RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research 5, 361-397(2004)

[13] Lewis, D. D. :Reuters-21,578 text categorization collection, http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html .(1999) accessed on 11 may 2013

[14] Lewis, D.D. and, and M. Ringuette: A Comparison of two learning algorithms for text categorization. In :Third Annual Symposium on Document Analysis and Information Retrieval (1994)

[15] Li,S. Z, Chan,K. L.:Performance Evaluation of The nearest feature line Method in Image Classification and Retrieval. IEEE Trans on Pattern Analysis and Machine Intelligence. 22(11),1335-1339(2000)

[16] Liu ,H., Sun J., Liu,L., Zhang, H.:Feature selection with dynamic mutual information. Pattern Recognition. 42 (7), 1330–1339(2009)

[17] Mladenic ,D., Grobelnik,M.:Feature selection on hierarchy of web documents. Decision Support Systems. 35 (1),45–87(2003)

[18] Ogura ,H., Amano,H., Kondo,M.:Feature selection with a measure of deviations from Poisson in text categorization.Decision Support Systems. 36 (3) ,6826–6832(2009)

[19] Parvin ,H., Alizadeh,H., Minaei,B.:Modified k Nearest Neighbor.Global Journal of Computer Science and Technology.10 (14),37-41(2010)

[20] Sebastiani, F. :Machine learning in automated text categorization.ACM Computing Surveys. 34(1),1–47(2002)

[21] Shang ,W., Huang,H., Zhu,H., Lin ,Y., Qu ,Y., Wang, Z.:A novel feature selection algorithm for text categorization.Expert Systems with Applications. 33 (1), 1–5(2007)

[22] Yang ,Y., Pedersen,J.O. :A comparative study on feature selection in text categorization.In: Proceedings of the 14th International Conference on Machine Learning(1997)

[23] Yong ,Z. :An Improved kNN Text Classification Algorithm based on Clustering. Journal of computers.4(3), 230–237(2009)

AUTHORS PROFILE

Siddhartha Kumar Arjaria has obtained the Bachelor of Engineering from Institute of technology, university of bilaspur(C.G.) in year 2001, and has obtained Master of technology from Pt. Ravi Shankar university,Raipur in 2006. Currently he is pursuing Ph.D in computer science & engineering departmentfrom Maulana Azad National Insitute of Technology ,Bhopal(M.P.)

Dr.Deepak singh Tomar has obtained his Ph.D in computer science & engineering from in Maulana Azad National Institute of Technology, Bhopal . Currently he is working as an Assistant Professor in Maulana Azad National Institute of Technology, Bhopal . His research area is Data mining . he has published about 3o papers in international & national journals.

D. Roy has obtained the Bachelor of Engineering in Electronics from Maulana Azad National Institute of Technology, Bhopal in the year 1990, Master of Engineering in Computer Science Engineering from National Institute of Technology, Rourkela in the year 1998 and Ph D in the year 2007 from the Department of Computer Science Engineering, Indian Institute of Technology, Kharagpur, India. She has worked with many prestigious institutes of India like Indian Institute of Technology, National Institute of Technology etc. in India. Currently she is working as an Associate Professor in Maulana Azad National Institute of Technology, Bhopal Dr. Roy has received a research grant of worth Rupees 9.73 lakhs from Government of India to carry out a research project. Total number of papers published in referred Journals, International conferences and International workshops are 25. Current research interest includes Information Retrieval, natural language processing and application of Artificial Engineering techniques in Electronic and mobile Learning.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Developing Extracting Association Rules System from Textual Documents

Arabi Keshk
Faculty of Computers and Information
Menoufia University
Shebin El-Kom, Egypt
arabikeshk@yahoo.com

Hany Mahgoub
Faculty of Computers and Information
Menoufia University
Shebin El-Kom, Egypt
h_mahgoub@yahoo.com

***Abstract*—A new algorithm is proposed for generating association rules based on concepts and it used a data structure of hash table for the mining process. The mathematical formula of weighting schema is presented for labeling the documents automatically and its named fuzzy weighting schema. The experiments are applied on a collection of scientific documents that selected from MEDLINE for breast cancer treatments and side effects. The performance of the proposed system is compared with the previous Apriori-concept system for the execution time and the evaluation of the extracted association rules. The results show that the number of extracted association rules in the proposed system is always less than that in Apriori-concept system. Moreover, the execution time of proposed system is much better than Apriori-concept system in all cases.**

Keywords- data mining; association rules; fuzzy system; apriori-concept system

I. INTRODUCTION

The explosive growth of information in textual documents creates a great need of techniques for knowledge discovery from text collections. Collecting, analyzing and extracting useful information from a very large amount of medical texts are difficult tasks for researchers in the medicine who need to keep up with scientific advances. Nowadays several domains in medical practice, drug development, and health care require support for such actives such as bioinformatics, medical informatics, clinical genomics, and many other sectors. Moreover, the examined textual data are generally unstructured as in the case of Medline abstracts in the available resources such as PubMed, search engine interfacing Medline and medical records. All these resources do not provide adequate mechanisms for retrieving the required information and analyzing very large amount of text content.

Text Mining is a tool to support and automate the process of finding and extracting interesting information from the documents. Selecting features are necessary and sufficient for constructing a model that can accurately predict future events or describe a problem. The models based on informative features will be easier to interpret from the other models, which are based on uninformative features. The quality of the features must be described in terms of semantic richness. For example, breast cancer is a disease occurring in a particular part of the body. If a text mining system represented this phrase using the two individual features breast and cancer, it would not capture the meaning of the phrase breast cancer. Thus, the concept feature breast cancer is semantically richer than the individual features breast and cancer. Therefore increasing the information content or semantic richness of the features will increase the plausibility and usefulness of the extracted association rules.

In this paper, we present a new text mining system that called developed extracting association rules from textual documents (D-EART) for extracting association rules from online structured and unstructured documents. The design of the D-EART system is based on concepts representation. D-EART is designed to overcome the drawbacks of the previous EART system that is presented in [1] and [2]. The mathematical weighting schema formula that used in the EART system is developed and is named fuzzy weighting schema. In addition, generation association rules based concept algorithm (GARC) is used for the mining process instead of word based as in the traditional data mining algorithms. In the D-EART system, MEDLINE abstracts are selected for the breast cancer treatments and side effects as the main domain of online collecting documents. The system is consists of three phases that are Text Preprocessing, Association Rule Mining (ARM), and visualization.

The reset of this paper is organized as follows. Section II presents the related work. Section III presents the D-EART system architecture. Experimental results are presented in section IV. Section V provides conclusion and future work.

II. RELATED WORK

There are several previous works in the field of association rules mining from structured documents (XML data) [3, 4, 5, 6 and 7]. More precisely the ability to extract useful knowledge from XML data is needed because the numerous data have been represented and exchanged by XML. Thought there are some works to exploit XML within the knowledge discovery tasks, and most of them rely on legacy relational database with an XML interface. In addition, mining knowledge in XML world is faced with more challenges than in the traditional well-structured world because of the inherent flexibility of XML. Extracting association rules from native XML documents called "XML association rules" was first introduced by Braga et al in [4]. All the previous works in this

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

field are based on the word features or structured data, consequently all extracted association rules are the relations between words [6, 7].

Recently, some works developed tools for extracting association rules from XML documents [8, 9], but both of them are approaching from the view point of XML query language. This caused the problem of language-dependent association rules mining. Ding et. al in [5] developed a method to discover all of the possible rules, i.e. generalized association rules from XML documents. In this method, all of the possible combinations of XML nodes based on their multiple nesting are used to generate the relational transactions format. This method suffered from some shortcomings such as generation of redundant rules. Moreover, it ignored the valuable tree structure of the documents.

A model for the effective extraction of generalized association rules from a collection of XML document is presented in [3]. This method does not used frequent subtree mining techniques in the discovery process and not ignored the tree structure of data in the final rules. The frequent subtrees based on the user that provide support and split to complement subtrees to form the rules. From the above previous works, we found that all works concentrated on the domain of Association Rules Mining (ARM) based on words from XML data documents. Therefore this research is concentrated on mining of association rules based on concepts from native XML text documents and deals with their tags.

In the field of ARM from unstructured documents, there is a large body of previous works. Identifying informative features from natural language (text) can be difficult so that the problem is that there are many approaches use semantically poor features, such as words [10]. These approaches take bag of words as input to the association rule mining algorithm such as Apriori algorithm, and finds associations among single isolated words. These approaches have the advantage of domain independent and easy to implement. There are two drawbacks in these approaches. Firstly, some concepts consist of multiple words, these multiple words concepts cannot be found as a unit in the association rules, and secondly the number of association rules is tremendously large.

There are some approaches was concentrated on extracted association rules based on concepts instead of words as in [11, 12, and 13]. The identified problems in these approaches are:

1) The ambiguity of the language and can be overcome with human interaction.

2) They used the Apriori algorithm to generate association rules based on concepts.

3) There are many systems based on word features representation and do not take into account the synonymy problem. These systems could cause a text mining system to generate a misleading model of association rules.

The earlier work of association rules mining from text has explored the use of manually assigned keywords [14]. They used keywords as features for generating association rules. The drawbacks of these approaches are that:

1) It is time consuming to manually assign the keywords.

2) The keywords are fixed (i.e., they do not change over the time or based on a particular user).

3) As the keywords are manually assigned, they are subject to discrepancy.

4) The textual resources are constrained to only those that have keywords.

Therefore, the work is needed to automate indexing of the textual document in order to allow the use of association extraction techniques on a large scale. Another research has been focused on constructing techniques to improve the quality of text-mined association rules. Most of these approaches generate a set of rules, and apply ranking techniques such as interestingness as in [15, 16].

Unlike these approaches, this research is focused on extracted the interesting set of the association rules. That rules are based on semantically richer representations. In mining area, most of previous studies adopt an Apriori for candidate set generation and test approach. However, candidate set generation is still costly, especially when there are a large number of patterns and/or long patterns [17]. Agrawal et al. had first introduced the problem of association rules mining [18]. Methods for association rules mining from both structured and unstructured documents have been well developed. Apriori and AprioriTid Algorithms are presented in [19]. These Algorithms, which are used for discovering large item sets make multiple passes over the data. This is the main problem of the Apriori algorithm since it reduces the performance of the system by increasing the time and generating tremendously large association rules where most of them are not plausible and useful.

III. D-EART SYSTEM ARCHITECTURE

The D-EART system is automatically discovers association rules from the collection of online structured and unstructured documents as shown in Fig. 1. It is designed to discover three types of relations such as:

1) The association rules amongst concepts only.

2) The association rules amongst the words only that are remained in the documents after extracted the concepts.

3) Get the relations between the concepts and words in the form of complex rules.

The modifications in the D-EART that overcome the drawbacks of the previous EART system in [1, 2] are as follows:

- On-line documents collecting and it accepts all native XML documents. The system designed for concepts representation, and it takes into account the characteristics of the natural language such as synonymy.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

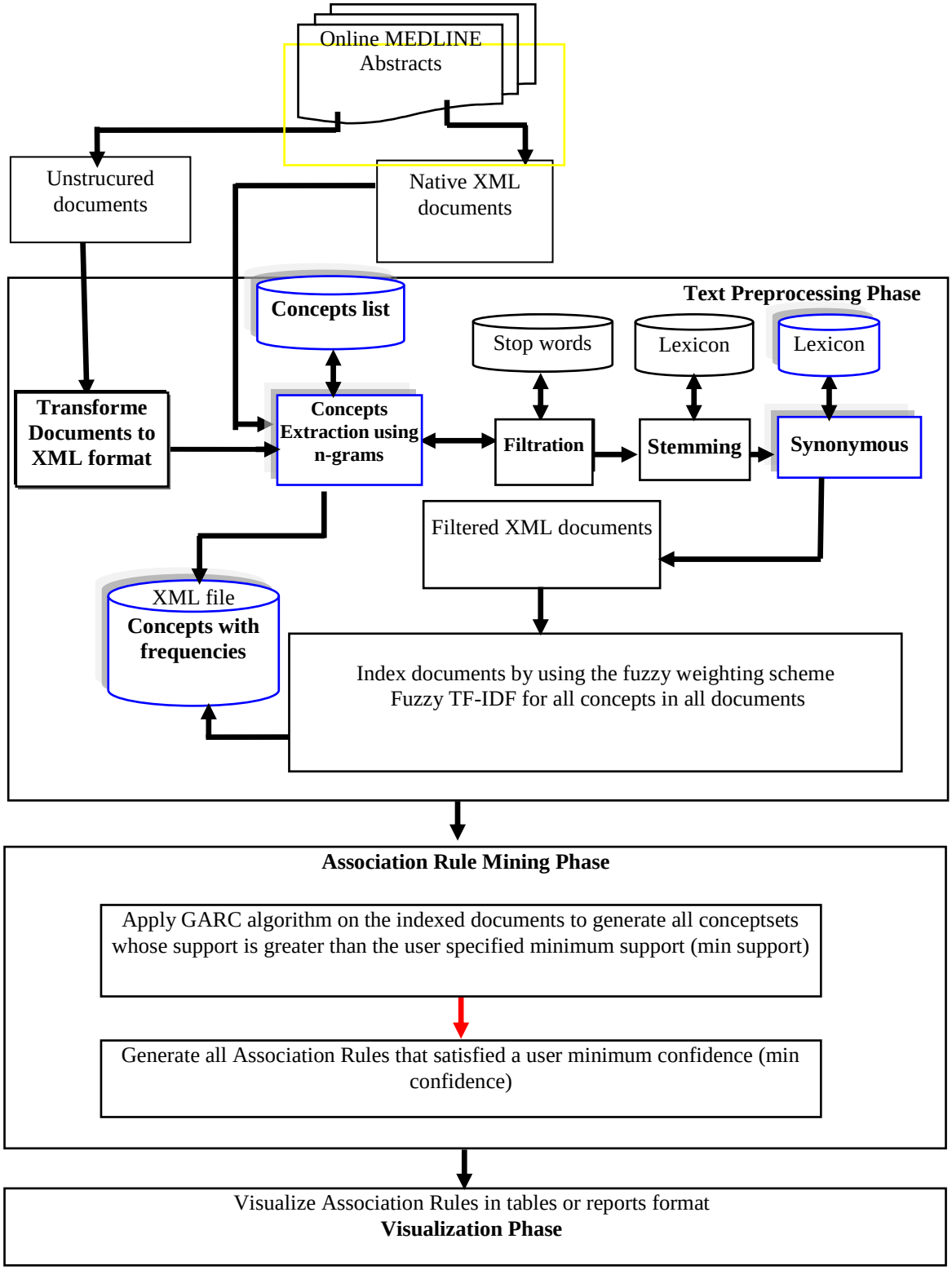


Figure 1. The D-EART system architecture

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- The system automatically indexing documents by using the developed fuzzy weighting schema without using the threshold weight value.
- The system designed based on a new algorithm for extracting association rules based on concepts (GARC). The algorithm overcomes the drawbacks of the previous algorithms by employing the power of data structure called hash table. Furthermore the system has the ability to perform different queries on the extracted association rules.

The D-EART system is consists of three main phases beside the online documents collection. The main phases are Text Preprocessing phase that include transformation, filtration, stemming, synonymy and indexing of documents, Association Rule Mining (ARM) phase that include a new GARC algorithm, and visualization phase.

A. *Online Documents Collection*

The D-EART system works online, so it is considered to be as a web-based text mining system. The D-EART accepts the documents that in XML format (structured) and also the unstructured documents. From the interface of the D-EART system, the user can online access the MEDLINE link and writes the search keywords. The selected documents and their tags are automatically loading into the system and the user selects the specific part of documents that will work on it.

B. *Text Preprocessing Phase*

The D-EART system has the ability to deal with the native XML documents and the unstructured documents. The process of concept extraction is done and the documents are filtered, stemmed and synonym used. Finally, the XML documents are automatically indexed by using the fuzzy weighting schema.

- *Transformation*

Once the online XML documents download into the system, their tags are automatically extracted in a combo box as shown in Fig. 2. The user can determine his specific part of the documents (for example the abstract part, </Abstract Text>) to work on it. Therefore the D-EART system is flexible to work on specific or all parts of documents. In the case of the unstructured documents, the D-EART system transforms them to the XML format.

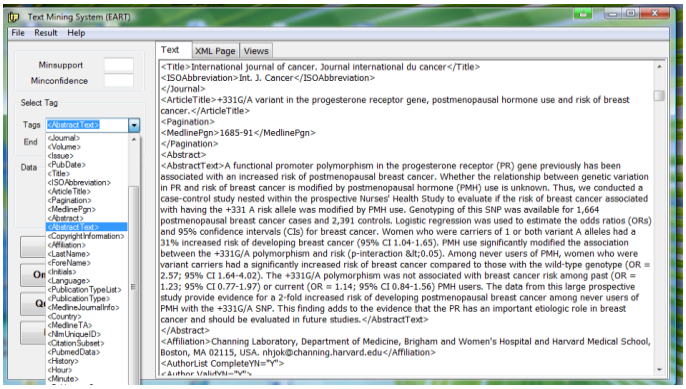


Figure 2. Selecting a specific tag of the documents

- *Concept Extraction*

The concept is a single word or a group of consecutive words that occurs frequently enough in the entire document collection. It is important to appear the concepts as a unit in the extracted association rules. The process of concept extraction as shown in Fig. 3 can be done as follows:

1) Splitting the documents into sentences by using the End-of-Sentence Detection Algorithm (ESDA) to determine the sentence boundary [20].

2) Determine each concept candidate using n-grams model [21]. We collect all the ordered pairs, or 2-grams, (A, B) such that words A and B occur in the same document in this order and the pair is frequent in the document collections.

3) Building a list of all concepts in the D-EART system, and map the concepts from concept list with sentences in documents and then estimate their frequencies.

4) Store all concepts with their frequencies in XML file.

1. For each document in the corpus
2. Sentence boundary←End of Sentence Detection Algorithm
3. Concept List
4. For each concept in the Concepts List
5. Count=0
6. For each sentence in the documents
7. N-grams concept in sentence ← Concept in Concepts List
8. Count +;
9. End for
10. End for
11. Concept File ← Each Concept with its frequencies

Figure 3. Concepts extraction process

- *Filtration*

The documents are filtered by removing the unimportant words from the documents. A list of unimportant words called stop words is built. The system checks the documents content and eliminates these unimportant words (e.g. articles, pronouns, conjunctions, and common adverbs). Moreover, the system replaces special characters, parentheses, commas, etc., with distance between words and concepts in the documents.

- *Stemming*

After the filtration process had done, the D-EART system automatically do word stemming based on the inflectional stemming algorithm which illustrated in [20]. The inflectional stemming algorithm consists of both part of rule-based and dictionary-based.

- *Synonymy*

In the synonymy process, the D-EART system matches each concept in the documents with the augment synonymous list. When the system finds a synonym for the concept, it replaces

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the concept in all documents with its synonymy. For example, the phrase hair loss is synonymous with the medical concept alopecia. The actual times occurs number of this concept is the total number of times that hair loss and alopecia occurs in the text. Since a concept representation would unify the expression hair loss with alopecia and thus account for synonymy. In contrast, the systems based on word representation would distribute the count between the three features hair, loss, and alopecia. The word based count would be smaller than the actual number of occurrences of the medical concept alopecia.

- *Indexing*

Mathematical formula of weighting schema in D-EART system is developed and used in [1, 2], and it named fuzzy weighting schema. This formula overcomes the drawbacks of the standard weighting schema. All weighted concepts are store in XML file for using them as input to the mining process.

- *The effect of Fuzzy Weighing Schema*

One of the drawbacks of the previous EART system is that the value of the threshold weight is hard. So we developed the system to automatically compute the weight value for each word and select the actually important concepts without entering the threshold weight value M. We developed the mathematical formula weighting schema and named it fuzzy weighting schema since the threshold weight value is replaced with the fuzzy membership value as shown in Equation (1)

$$\mu_{i,j} = \left\{ \frac{Nt_j}{|C|} \quad where \quad 0 \le \mu \le 1 \right. \tag{1}$$

Where Nt_j denotes the number of documents in collection C in which t_j occurs at least once (document frequency of the term t_j) and $|C|$ denotes the number of the documents in collection C. Therefore, the fuzzy weighting schema is defined as follows:

$$Fuzzy \cdot w(i,j) = \mu_{i,j} * \begin{cases} Nd_i, t_j * \log_2 \dfrac{|C|}{Nt_j} & if \ Nd_i, t_j \ge 1 \\ 0 & if \ Nd_i, t_j = 0 \end{cases} \tag{2}$$

This formula caused developing in the system since the high weighted values were given to the concepts that are more occurrences in the documents. Moreover, new concepts appeared with high fuzzy weighted values although they are disappeared by using the weighing schema. The D-EART system automatically eliminates 10% of all concepts that have low weighted values. After that the system stores all concepts without redundancy with their frequencies in XML file for using them as input to the mining process.

- *Fuzzy Weighting Schema Case Study*

Consider the 6-collection of documents as shown in Fig. 4. In the indexing process, the fuzzy weighted values are calculated for each concept in the 6 documents. The total number of concepts is equal to 21 concepts in all documents. We summarized all concepts with their two weighted values in Table I.

Collection of Documents	
DID	**Concepts**
D1	$C_1\ C_2\ C_1C_3\ C_6\ C_4$
D2	$C_3\ C_4\ C_5\ C_3\ C_5\ C_5\ C_4$
D3	$C_2\ C_3C_4\ C_2\ C_3\ C_3\ C_3\ C_5$
D4	$C_1C_5\ C_4\ C_1\ C_5\ C_1\ C_5\ C_5\ C_5$
D5	$C_3\ C_4\ C_5\ C_3\ C_4\ C_5\ C_3$
D6	$C_2\ C_5\ C_4\ C_5\ C_2\ C_5\ C_2\ C_5$

Figure 4. The collection with 6 documents.

TABLE I. THE TF-IDF AND FUZZY TF-IDF VALUES FOR EACH CONCEPT IN SIX DOCUMENTS

D-ID	Concept	Frequencies	No. of documents	TF-IDF	Fuzzy TF-IDF
D1	C_1	2	2	0.954	0.318
	C_2	1	3	0.301	0.151
	C_3	1	4	0.176	0.117
	C_6	**1**	**1**	**0.778**	**0.129**
	C_4	**1**	**6**	**0.0**	**0.0**
D2	C_3	2	4	0.352	0.235
	C_4	2	6	**0.0**	**0.0**
	C_5	3	5	0.237	0.197
D3	C_2	2	3	0.602	0.301
	C_3	4	4	0.704	0.469
	C_4	1	6	**0.0**	**0.0**
	C_5	1	5	0.079	0.066
D4	C_1	3	2	1.431	0.477
	C_4	**1**	**6**	**0.0**	**0.0**
	C_5	5	5	0.395	0.329
D5	C_3	3	4	0.528	0.352
	C_4	**2**	**6**	**0.0**	**0.0**
	C_5	2	5	0.158	0.132
D6	C_2	3	3	0.903	0.452
	C_4	**1**	**6**	**0.0**	**0.0**
	C_5	4	5	0.316	0.263

From Table I, it noticed that a concept C4 has zero weighted values so that the system automatically eliminates it from all documents. The system resorts the concepts based on their weighted values from the highest to the lowest.

Table II shows all resorted concepts with their TF-IDF values. By choosing the threshold weight value M=50%, all concepts that in the shaded region had discarded. The system stores all accepted concepts without redundancy which are approximately 4 concepts (C1, C2, C3 and C6) in XML file.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II
THE CONCEPTS WITH THEIR TF-IDF.

Concept	Documents	TF-IDF
C_1	D4	1.431
C_1	D1	0.954
C_2	D6	0.903
C_6	**D1**	**0.778**
C_3	**D3**	**0.704**
C_2	D3	0.602
C_3	D5	0.528
C_5	**D4**	**0.395**
C_3	D2	0.352
C_5	D6	0.316
C_2	D1	0.301
C_5	D2	0.237
C_3	D1	0.176
C_5	D5	0.158
C_5	D3	0.79

TABLE III
THE CONCEPTS WITH THEIR FUZZY TF-IDF.

Concept	Documents	Fuzzy TF-IDF
C_1	D4	0.477
C_3	**D3**	**0.469**
C_2	D6	0.452
C_3	D5	0.352
C_5	**D4**	**0.329**
C_1	D1	0.318
C_2	D3	0.301
C_5	D6	0.263
C_3	D2	0.235
C_5	D2	0.197
C_2	D1	0.151
C_5	D5	0.132
C_6	**D1**	**0.129**
C_3	D1	0.117
C_5	D3	0.066

Table III shows the same resorted concepts but with their Fuzzy TF-IDF values. The concepts that appear in the shaded region had discarded, since the less important concepts with fewer frequencies always exist in the bottom of the table. After that the system stores all concepts without redundancy with their frequencies which are approximately 4 concepts (C1, C2, C3 and C5) in XML file for using them as input in the mining process.

It noticed that the descending order of the concepts becomes different from the order in Table II. The main reasons for the difference are:

1) The first effect of the fuzzy weighting schema, since the high weighted values are given to the concepts that are more occurrences in documents. For example, the concept C6 is in two different orders as shown in Table II and III. The weighing schema considered the concept C6 an important concept although it occurred only one time in all documents.

2) The second effect of the fuzzy weighting schema is the appearing of new concepts with high fuzzy weighted values in the top of the list. For example, in Table 2 the concept C5 does not satisfy the threshold weight value although C5 occurred 5 times in D4. In contrast in Table III the concept C5 has a high fuzzy weighted value and exists in the top of the table.

C. Association Rule Mining (ARM) Phase

The D-EART system designed to extract association rules based on concepts by using a new GARC algorithm. The algorithm overcomes the drawbacks of the Apriori algorithm by employing the power of data structure called Hash Table. The hashing function h(v) and concepts number (N) considered the key factors in hash table building and search performance. The GARC algorithm is utilized with dynamic hash table.

- *Generating Association Rules Algorithm based on Concepts (GARC)*

The proposed GARC algorithm as in Fig. 5 employs the following two main steps:

1) Based on the number of concepts N in the documents, a dictionary table was constructed as shown in Table IV for N = 6 concepts.

2) There are also two main processes for a dynamic hash table: the building process, and the scanning process. The mining process for GARC algorithm includes the two processes (building and scanning process) on the given XML file that contains all concepts.

The hash function h(v) = v mod N, where v is a key (location of primary concept) is used to build a primary bucket of the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

hash table. The algorithm scans only the XML file that contained all important concepts not the original documents. The scanning process is done as follows:

1) Make all possible combinations of concepts then determine their locations in the dynamic hash table by using the hash function h(v).

2) Insert all concepts and concept sets in a hash table and update their frequencies, the process continues until there is no concept in the XML file.

3) Save the dynamic hash table into secondary storage media.

4) Scan the dynamic hash table to determine the large frequent concept sets that satisfy the threshold support.

- *Advantages of GARC algorithm*

The advantages of the GARC algorithm summarize as follows:

1) The algorithm permits the end user to change the threshold support and confidence factor,

2) Small size of dynamic hash table, since with changing the size of concept set the size of dynamic hash table will change.

3) Less number of concept sets, since there is no concept sets with zero occurrences will occupy a size in a dynamic hash table.

GARC_algorithm ( )

1. **Input** minimum support (*s*), minimum confidence (c), the number of concepts (*N*)
2. Build a primary bucket of hash table
3. **IF** there is no EOF **THEN**
4. **FOR** each document *D* (*d*(1) *d*(2) ... *d*(*n*)) **DO**
5. Select each concept *c*(1), *c*(2), ..., *c*(*N*)
6. Create all combinations of conceptset with their occurrences
7. Insert all conceptsets with their occurrences in hash table by using h(*v*)
8. **IF** there is document *D* **THEN**
9. Goto line 4
10. **ELSE**
11. Goto line 17
12. **ENDIF**
13. **ENDFOR**
14. **ELSE**
15. Goto line 19
16. **ENDIF**
17. Determine all large frequent conceptsets that satisfies the minimum support
18. Extract all Association Rules that satisfies minimum confidence
19. **STOP**

Figure 5. GARC algorithm

- *GARC algorithm Case Study*

The D-EART system run on a collection of 100 online XML documents selected from MEDLINE by thresholds values: support s=2% and confidence c=50%. The number of concepts *N*= 30 resulted from the indexing process and used for building a dynamic hash table. Fig. 6 shows the number of all fuzzy weighted concepts that labeling each document. Fig. 7 shows the number of the resultant association rules with c = 50%which is equal to 64 rules.

The D-EART system can do different queries on the extracted association rules. The query supports the medical researchers by a model of important relationships within the concept features. This model might identify relations between the disease and the suitable treatments, and relations between a treatment and its side effects. Fig. 8 shows the query screen which includes both the categories information and the queries result icons. The user can determine which the categories will get the relations between them. The query results can be saved on the hard disk through the export icon.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

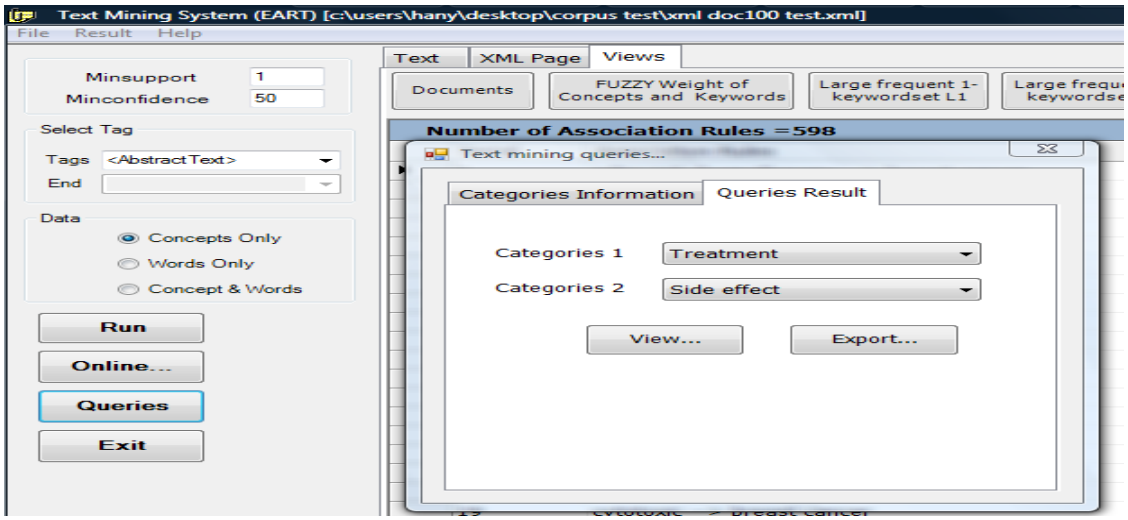
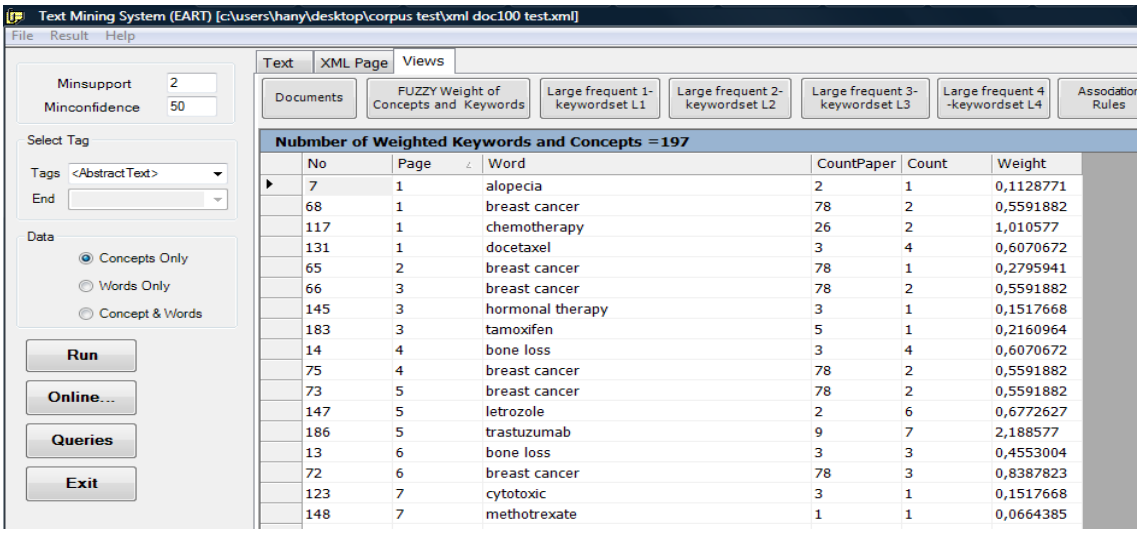
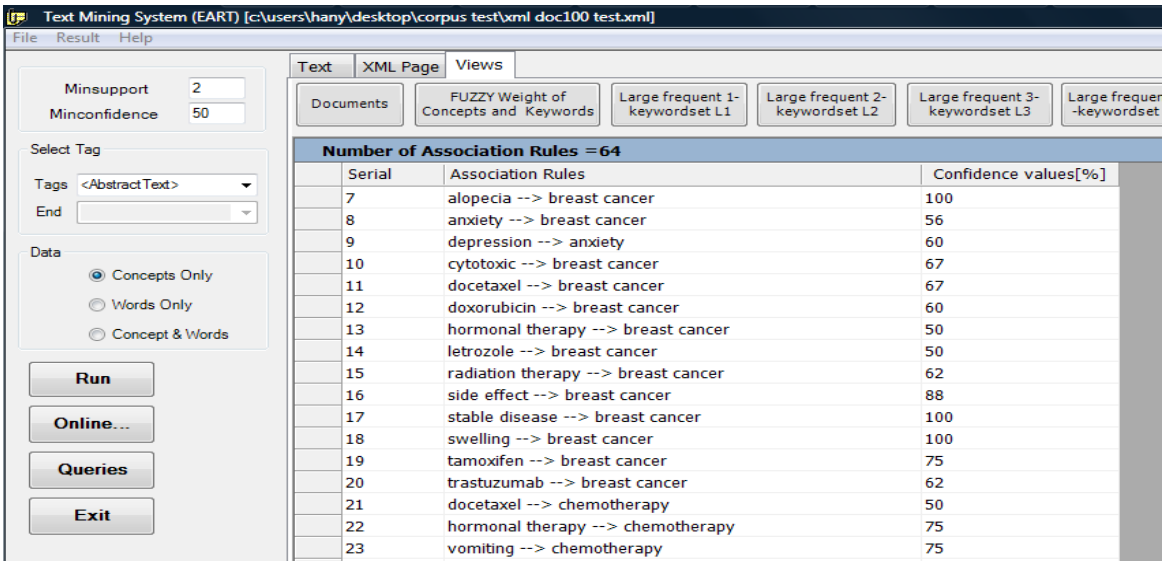


Figure 6. The number of fuzzy weighted concepts



No	Page	Word	CountPaper	Count	Weight
7	1	alopecia	2	1	0,1128771
68	1	breast cancer	78	2	0,5591882
117	1	chemotherapy	26	2	1,010577
131	1	docetaxel	3	4	0,6070672
65	2	breast cancer	78	1	0,2795941
66	3	breast cancer	78	2	0,5591882
145	3	hormonal therapy	3	1	0,1517668
183	3	tamoxifen	5	1	0,2160964
14	4	bone loss	3	4	0,6070672
75	4	breast cancer	78	2	0,5591882
73	5	breast cancer	78	2	0,5591882
147	5	letrozole	2	6	0,6772627
186	5	trastuzumab	9	7	2,188577
13	6	bone loss	3	3	0,4553004
72	6	breast cancer	78	3	0,8387823
123	7	cytotoxic	3	1	0,1517668
148	7	methotrexate	1	1	0,0664385

Figure 7: The resultant rules that satisfy s = 2%, c = 50% for Document=100, N=30.



Serial	Association Rules	Confidence values[%]
7	alopecia --> breast cancer	100
8	anxiety --> breast cancer	56
9	depression --> anxiety	60
10	cytotoxic --> breast cancer	67
11	docetaxel --> breast cancer	67
12	doxorubicin --> breast cancer	60
13	hormonal therapy --> breast cancer	50
14	letrozole --> breast cancer	50
15	radiation therapy --> breast cancer	62
16	side effect --> breast cancer	88
17	stable disease --> breast cancer	100
18	swelling --> breast cancer	100
19	tamoxifen --> breast cancer	75
20	trastuzumab --> breast cancer	62
21	docetaxel --> chemotherapy	50
22	hormonal therapy --> chemotherapy	75
23	vomiting --> chemotherapy	75

Figure 8. Query screen

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The advantages of D-EART system are as follows:

1) The user can access XML textual documents online.

2) The design of the D-EART system is based on concept representation and considers the synonymy as a characteristic of the natural language characteristics.

3) It is flexible to work on specific or all parts of the documents with the same structure. Moreover it is not fully domain-independent so we can apply it on other domains.

4) The proposed GARC algorithm overcomes the drawbacks of the previous algorithms.

5) It extracts three types of the association rules depending on the analysis of relations between the concepts only, words only and concepts with words. In addition different queries are available on the extracted association rules.

IV. EXPERENTAL RESULTS

The experiments are performed to compare the performance of both D-EART system and Apriori-concept system for the number of extracted association rules and the execution time. Finally, evaluate the performance of D-EART system at the three semantic levels: concepts only, words only, and concepts with words. The corpus of the PubMed abstracts that used in the experiments is consists of 10000 biomedical abstracts with keyword search "breast cancer treatments and side effects" [22]. All experiments are applied on the 10000 documents after divided them into six document sets 50, 100, 500, 1000, 5000, and all 10000 documents. The systems are implemented by using VS .Net 2005 (C#) and the experiments were performed on Intel Core2 Duo, 1.8 GHz system with Windows XP and 2 Giga of RAM.

A large number of association rules can be extracted by selecting the values of minimum support and confidence in the mining process. The D-EART system gives the best results by using low support and high confidence values. Moreover, the number of concepts that entered to the mining process is fewer by using the fuzzy weighting schema.

Table V shows the experiments that are applied on various documentsets by different threshold values. It noticed that the number of extracted association rules in D-EART system is useful and always less than that in Apriori-concept system. The reason returns to the strong effect of using the fuzzy weighting schema in D-EART system.

Fig. 9 (a) and Fig. 9 (b) show that the execution time of Apriori-concept system is increased regularly when the document sets are increased compared to D-EART system. The mining process in Apriori-word system takes more time for less number of concepts in the documents. The reason is that the mining process in Apriori algorithm depends on the size of documents rather than the number of concepts. The results show that the execution time of Apriori-concept system is about seventh fold of D-EART system. The D-EART system scans the documents only one time as the number of documents increased. Therefore the size of documents does not influence in the mining process. Finally, the results reveal that the execution time for D-EART system is much better than that of the Apriori-concept system in all cases.

TABLE V. THE NUMBER OF ASSOCIATION RULES FOR DIFFERENT THRESHOLD VALUES

No. of Documents	**Systems**	**Minimum Support (*s*), Minimum Confidence (*c*)**			
		s =1%, *c* =50%	3% 50%	7%, 60%	10%, 50%
500	Apriori-concept based	183	76	17	10
	D-EART	71	31	5	2
1000	Apriori-concept based	227	91	11	8
	D-EART	86	34	4	3
5000	Apriori-concept based	**239**	75	20	15
	D-EART	**92**	27	4	2
10000	Apriori-concept based	345	102	37	30
	D-EART	135	39	10	7

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

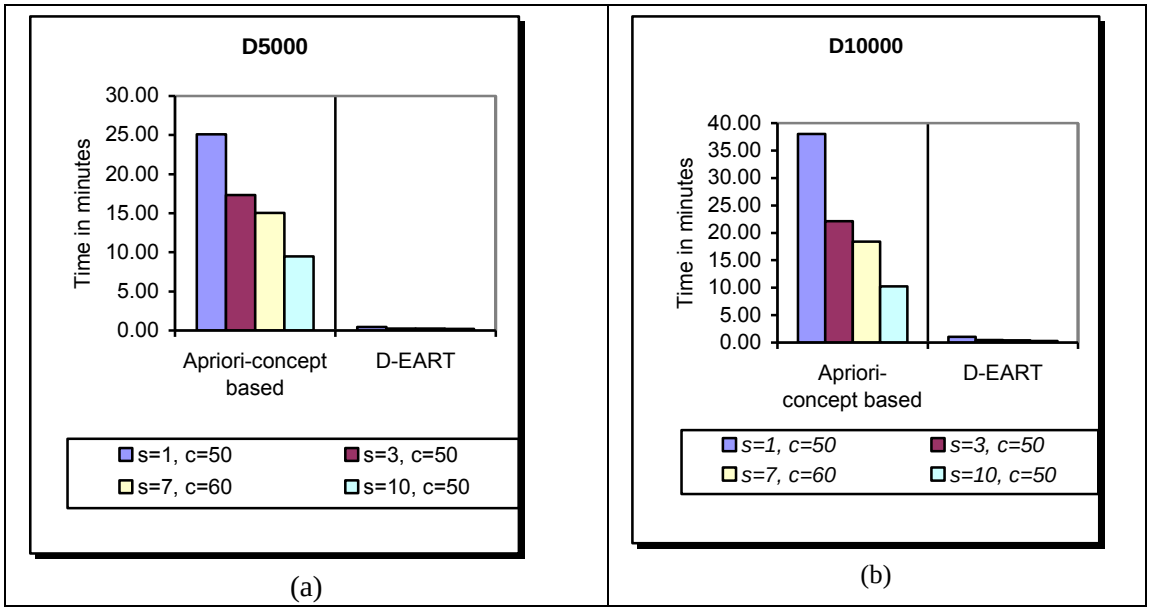


Figure 9. Execution time of Apriori-concept and D-EART systems

V. Conclusion and Future Work

This paper presented a new text mining system for extracting association rules based on concepts representation from online textual documents. This system overcame some of the problems in the previous EART system and the drawbacks of the Apriori algorithm by using the data structure hash table in the mining process. The results of comparing D-EART and Apriori-concept systems reveal that the number of extracted association rules in D-EART system is always less than that in Apriori-concept system. Moreover, the execution time for D-EART system is much better than that of Apriori-concept system in all cases. So concept technique would be suitable to apply to any large corpus of medical text such as portions of the web.

In future work we intend to apply D-EART system on PDF full text document with figures and images instead of using only the abstract part of documents.

References

[1] H. Mahgoub and D. Rösner "Mining association rules from unstructured documents", In Proc. 3rd Int. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic, Aug. 25-27, 2006, pp. 167-172.

[2] H. Mahgoub, D. Rösner, N. Ismail and F. Torkey, "A Text Mining Technique Using Association Rules Extraction", Int. J. of Computational Intelligence, Vol.4, Nr.1 2007 WASET.

[3] R. AliMohammadzadeh, M. Rahgozar, and A. Zarnani, "A New model for discovering XML association rules from XML Documents", In Proc. 3rd Int. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic, Aug. 25-27, pp. 365-369, 2006.

[4] D. Braga, A. Campi, M. Klemettinen, and P. L. Lanzi, "Mining association rules from XML data", In Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, September 4-6, Aixen-Provence, France 2002.

[5] Q. Ding, K. Ricords, J. Lumpkin, "Deriving General Association Rules from XML Data", In Proceedings of Fourth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'03) October 16-18, 2003 Lübeck, Germany.

[6] J. Paik, H. Yong Youn, and U. Kim, "A New Method for Mining Association Rules from a Collection of XML Documents", ICCSA 2005, LNCS 3481, pp. 936–945, 2005. Springer-Verlag Berlin Heidelberg 2005.

[7] J. Shin, J. Paik, and U. Kim, "Mining Association Rules from a Collection of XML Documents using Cross Filtering Algorithm", International Conference on Hybrid Information Technology (ICHIT'06) IEEE, 2006.

[8] D. Braga, A. Campi, S. Ceri, M. Klemettinen, and P. L. Lanzi, "A tool for extracting XML association rules", In Proc. of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02), pp.57–64, 2002.

[9] J. W. W. Wan and G. Dobbie, "Extracting association rules from XML documents using XQuery", In Proc. of the 5th ACM International Workshop on Web Information and Data Management (WIDM'03), pp.94–97, 2003.

[10] G. Paynter, I. Witten, S. Cunningham, and G. Buchanan, "Scalable browsing for large collections: a case study", 5th Conf. digital Libraries, Texas, pp.215-218, 2000.

[11] W. Jin, R. K. Srihari, and X. Wu, "Mining Concept Associations for Knowledge Discovery Through Concept Chain Queries", Z.-H. Zhou, H. Li, and Q. Yang (Eds.): PAKDD 2007, LNAI 4426, pp. 555–562, 2007.Springer-Verlag Berlin Heidelberg 2007.

[12] H. Murfi and K. Obermayer, "A Two-Level Learning Hierarchy of Concept Based Keyword Extraction for Tag Recommendations", Available at http://www.kde.cs.unikassel.de/ws/dc09/papers/paper_17.pdf, 2009.

[13] M. Roche, J´erˆome Az´e, O. Matte-Tailliez, and Y. Kodratoff, "Mining texts by association rules discovery in a technical corpus", Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[14] A. Amir, Y. Aumann, R. Feldman, and M. Fresko, "Maximal association rules: A tool for mining Associations in text", Journal of Intelligent Information Systems, 25:3, pp. 333-345, 2005.

[15] P. Feng H. Zhang Q. Qiu and Z. Wang, "PCAR : an Efficient Approach for Mining Association Rules", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE 2008.

[16] Y. Liu, S. Navathe, A. Pivoshenko, A. Dasigi, R. Dingledine and B. Ciliax, "Text analysis of Medline for discovering functional relationships among genes: evaluation of keyword extraction weighting schemes", Int. J. Data Mining and Bioinformatics, Vol. 1, No 1, 2006.

[17] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", In W.Chen, J. Naughton, and P. A. Bernstein, editors, 2000 ACM SIGMOD Intl. Conference on Management of Data, pp. 1-12. ACM Press, 05 2000.

[18] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", In Buneman, Peter and Jajodia, Sushil (Eds.), Proceedings of the 1993 ACMSIGMOD International Conference on Management of Data, Washington, D.C., pp. 207–216, 1993.

[19] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proc. 20th Int. conf. of very Large Data Bases, VLDB, Santigo, Chile, pp. 487-499, 1994.

[20] S. Weiss, N. Indurkhya, T. Zhang and F. Damerau, TEXT MINING: Predictive Methods for Analyzing Unstructured Information. Springer Science-business Media, Inc. 2005.

[21] P. Majumder, M. Mitra and B. Chaudhuri, "N-gram: a language independent approach to IR and NLP", International Conference on Universal Knowledge and Language (ICUKL), Goa, India, November, 2012.

[22] Available via the NCBI (the U.S. National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/) Entrez retrieval system: http://www.ncbi.nlm.nih.gov/pubmed

A novel congestion control mechanism for Traffic management in wireless sensor networks

Nasrin Azizi[1], Solmaz Abdollahi Zad[2]
Department of Computer, College of Computer, Sardroud Branch, Islamic Azad University, Sardroud, Iran
[1]nasrin_azizi61@yahoo.com
[2]Sk_abdolahi@yahoo.com

Abstract- **Due to the nature of wireless sensor networks the higher amount of traffic is observed when the monitored event takes place. Exactly at this instance, there is a higher probability of congestion appearance in the network. Congestion can cause missing packets, low energy efficiency, and long delay. Moreover, some applications, e.g. multimedia and image, need to transmit large volumes of data concurrently from several sensors. These applications have different delay and QoS requirements. Congestion problem is more urgent in such applications. Therefore congestion in WSNs needs to be controlled for high energy-efficiency, to prolong system lifetime, improve fairness, and improve quality of service in terms of throughput and packet loss ratio along with the packet delay. To achieve this objective, a novel congestion control protocol for traffic management is proposed in this paper. Proposed protocol can control congestion in the node and adjusts every upstream traffic rate with its node dynamic priority to mitigate congestion. Proposed protocol can broadcast traffic on the entire network fairly. Simulation results show that the performance of proposed protocol is more efficient than previous algorithms in terms of throughput.**

Keywords-wireless sensor network; congestion mitigation; traffic distribution; throughput

I. INTRODUCTION

A wireless sensor network (WSN) is a network composed of many sensor nodes capable of sensing a phenomenon, transforming the analog data to digital and transmitting them to destination nodes (usually called sinks). Due to severe power constraints their computation capability, as well as their transmission range, are limited. Thus, for the transmission of data from a source (the node that sensed the phenomenon) to a sink (the end-node that receives the data), the wireless sensor nodes that lie over between them, form a ‘‘path’’ and data are transmitted through them in a hop-by-hop manner.

However, with the specific consideration of the unique properties of sensor networks such limited power, stringent bandwidth, dynamic topology (due to nodes failures or even physical mobility), high network density and large scale deployments have caused many challenges in the design and management of sensor networks. These challenges have demanded energy awareness and robust protocol designs at all layers of the networking protocol stack [1].

The upstream traffic from sensor nodes to the sink is many-to-one multi-hop convergent. The upstream traffic can be classified into four delivery models: event-based, continuous, query-based, and hybrid delivery. Due to the convergent nature of upstream traffic, congestion more probably appears in the upstream direction. Congestion that can leads to packet losses and increased transmission latency has direct impact on energy-efficiency and application QoS, and therefore must be efficiently controlled. Congestion control generally follows three steps: congestion detection, congestion notification, and rate-adjusting.

In response to congestion, a rate adjustment mechanism must be designed and implemented properly in order to eliminate or avoid congestion. A number of different schemes were reported in the literature in last few years. A stop-and-start and hop-by-hop strategy is employed in [2]. In [4] and [7], an end-to-end and AIMD-like (Additive Increase Multiplicative Decrease) rate adjustment approach is employed. All of these mechanisms, however, aim at guaranteeing the simple fairness instead of the weighted fairness. A priority-based congestion control protocol (PCCP) is presented to achieve the weighted-fairness transmission for single-path routing WSNs in [10].

Two types of congestion could occur in WSNs [5]. The first type is node-level congestion that is common in conventional networks. It is caused by buffer overflow in the node and can result in packet loss, and increased queuing delay. Packet loss in turn can lead to retransmission and therefore consumes additional energy. For WSNs where wireless channels are shared by several nodes using CSMA like (Carrier Sense Multiple Access) protocols, collisions could occur when multiple active sensor nodes try to seize the channel at the same time. This can be referred to as link level congestion. Link-level congestion increases packet service time, and decreases both link utilization and overall throughput, and wastes energy at the sensor nodes. Both node-level and link-level congestions have direct impact on energy-efficiency and QoS.

Congestion control protocol efficiency depends on how much it can achieve the following performance objectives: (i) First, energy-efficiency requires to be improved in order

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

to extend system lifetime. Therefore congestion control protocols need to avoid or reduce packet loss due to buffer overflow, and remain lower control overhead that will consume additional energy more or less. (ii) Second, fairness needs to be observed so that each node can achieve fair throughput. Fairness can be achieved through rate-adjustment and packet scheduling (otherwise referred to as queue management) at each sensor node. (iii) Furthermore, support of traditional quality of service (QoS) metrics such as packet loss ratio and packet delay along with throughput may also be necessary.

The rest of the paper is organized as follows: in section 2, we explain the related works. Section 3 describes the System Models and Proposed Protocol Discussion. Section 4 describes simulation parameters and result analysis. Final section is containing of conclusion.

II. RELATED WORKS

There are several congestion control protocols [3], [5]-[8], [12, 13] for sensor networks. They differ in the way that they detect congestion, broadcast congestion related information, and the way that they adjust traffic rate when congestion occurs. In this section, we review some of them and discuss their limitations.

Congestion detection and avoidance (CODA) [7] proposes an open-loop, hop-by-hop backpressure mechanism and a closed-loop, multi-source regulation mechanism in event-driven

WSNs. Sensor nodes detect congestion by monitoring the channel utilization and buffer-occupancy level. In response to congestion, the congested sensor nodes send backpressure messages to their neighbors which may drop packets, reduce their sending rate and further propagate backpressure messages. If the sending rate of a source node is greater than the preset threshold, the source node must receive a continuous stream of ACKs from the base station in order to maintain that rate. By this means, the base station may limit the sending rate of a source node based on deciding how many ACKs to broadcast. CODA employs the AIMD (Additive Increase Multiplicative Decrease) coarse rate adjustment. It only guarantees simple fairness of the congestion control.

Event-to-sink reliable transport protocol (ESRT) [9] monitors the local buffer level in intermediate sensor nodes and sets a congestion notification bit in the packet when the buffer overflows. If a base station receives a packet whose congestion notification bit is set, it broadcasts a control signal to inform all source nodes to reduce the sending rate according to certain proportion. ESRT limits sending rate of all source nodes when congestion occurs regardless of where the hot spot happens in WSNs. The best way is to regulate those source nodes that are responsible for this congestion.

Priority based congestion control protocol (PCCP) [10] defines a new variable, congestion degree as ratio of average packet service time over average packet inter-arrival time at each sensor node. Congestion degree is intended to reflect the current congestion level of each sensor node. Based on congestion degree, PCCP employs a hop-by-hop rate adjustment technique called priority-based rate adjustment (PRA) to adjust the scheduling rate and the source rate of each sensor node in a single-path routing WSN. In the tree-based network topology of single-path routing WSNs, a sensor node will only has one downstream neighbor, but it may have multiple upstream neighbors. The whole data flow generated by a source node will pass through the nodes and links along with the single routing path. Sensor nodes learn the number of upstream data sources in the sub tree roots and measure the maximum downstream forwarding rate. Finally, they calculate the per-source rate based on priority index of each source node.

In Fusion [11], congestion is detected in each sensor node based on measurement of queue length. The node that detects congestion sets a CN (congestion notification) bit in the header of each outgoing packet. Once the CN bit is set, neighboring nodes can overhear it and stop forwarding packets to the congested node so that it can drain the backlogged packets. This non-smooth rate adjustment could impair link utilization as well as fairness, although Fusion has a mechanism to limit the source traffic rate and a prioritized MAC algorithm to improve fairness.

Adaptive Rate Control (ARC), [8], is an LIMD-like (linear increase and multiplicative decrease) algorithm. In ARC, if an intermediate node overhears that the packets it sent previously are successfully forwarded again by its parent node, it will increase its rate by a constant α. Otherwise it will multiply its rate by a factor β where $0 < \beta < 1$. ARC does not use explicit congestion detection or explicit congestion notification and therefore avoids use of control messages. However the coarse rate adjustment could result in tardy control and introduce packet loss.

CCF (Congestion Control and Fairness) [5] uses packet service time to deduce the available service rate and therefore detects congestion in each intermediate sensor node. Congestion information, that is packet service time in CCF, is implicitly reported. CCF controls congestion in a hop-by-hop manner and each node uses exact rate adjustment based on its available service rate and child node number. CCF guarantees simple fairness. That means each node receives the same throughput. However the rate adjustment in CCF relies only on packet service time which could lead to low utilization when some sensor nodes do not have enough traffic or there is a significant packet error rate (PER).

Those existing congestion control protocols for WSNs have limitations. For example, they only guarantee simple fairness, which means that the sink receives the same throughput from all nodes. However, sensor nodes may have different priority or importance due to either their functions or the location at which they are deployed. Also they work on single-path routing WSNs.

III. SYSTEM MODELS AND PROPOSED PROTOCOL DISCUSSION

This section describes network and node models, as shown in Figs. 1 and 3, respectively. Also this section explains the proposed protocol with details.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Network Model

The network model to be investigated in this work is depicted in Fig. 1, where sensor nodes are supposed to generate continuous data and form many-to-one convergent traffic in the upstream direction. Data packets are sent from nodes to sink through a multi-path and multi hop network. CSMA/CA MAC protocol is implemented in MAC layer. Each sensor node could have two types of traffic: source and transit. The source traffic is locally generated at each sensor node, while the transit traffic is from other nodes. As shown in Fig. 1, node 1 is a source node and only has source traffic, while nodes 2, 3 and 4 are source nodes as well as intermediate nodes because they have source traffic as well as transit traffic.

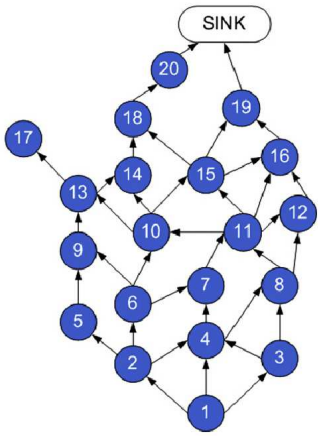


Figure 1. General network model

Every packet has two kinds of priorities: static priority and dynamic priority, which will be defined shortly. Packet static priority is represented as an integer and the lowest static priority SP (packet) = 0. Dynamic priority changes with the number of spent hops as follows:

$$DP = \frac{hop_s}{sink\_level} + SP \quad (1)$$

B. Level discovery

This phase each node that is becoming a source node is self-assigned as level 0 and sends a *level_discovery* message to the neighbors selected during topology control phase. Nodes that receive this packet are considered as children to the source node and are set as level 1. Each of these nodes broadcast again the *level_discovery* packet, and the pattern continues with the level 2 nodes. This procedure iterates until all nodes are assigned a level and stops when the *level_discovery* packets reach the sink.

The hierarchical tree algorithm is also able to identify and rectifysome issues that are possible to arise. Specifically, it is possible for a node to be the last one that receives the level_discovery packet when there are no other nodes upstream able to forward that packet. This node responds by broadcasting a ‘‘negative ACK’’ packet (NACK) indicating that it cannot route any packets. When the upstream nodes receive the NACK they become aware of the situation.

An example of the operation of hierarchical tree algorithm, and placement of nodes in levels, is illustrated in Fig. 2.

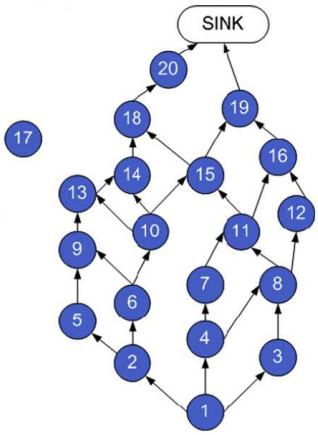


Figure 2. Network model after level discovery

C. Node Model

Node model of the investigated wireless sensor network is presented in Fig. 3. The source traffic of node *i* is generated with source traffic rate (r_s^i) by itself locally. The transit traffic of node *i* is received with transit traffic rate (r_{tr}^i) from its child nodes through MAC layer of node *i*. Both r_s^i and r_{tr}^i are converged through network layer to MAC layer as total input rate of node *i* (r_{in}^i).

If buffer incoming rate is r_{in}^i and buffer outgoing rate is r_{out}^i, then buffer changing rate $R = r_{out}^i - r_{in}^i$ and weighted buffer changing rate is

$$WR = DP(r_{out}^i) - DP(r_{in}^i) \quad (2)$$

Weighted queue length is defined as

$$WQ = \sum_{i=1}^{N} DP(packet_i) \quad (3)$$

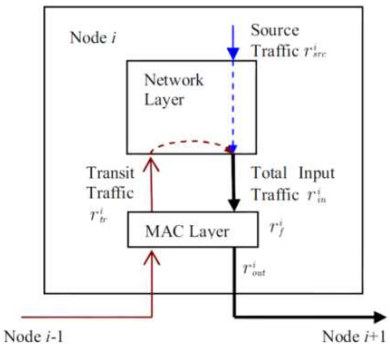


Figure 3. General node model

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

D. Congestion Control Mechansim

In order to precisely measure local congestion level at each node, we propose dual buffer thresholds and weighted buffer difference for congestion detection. Buffer is defined as three states, “accept state”, “filter state” and “reject state”, as Fig. 4 indicate. Two thresholds Q_{min} and Q_{max} are used to border different buffer states.

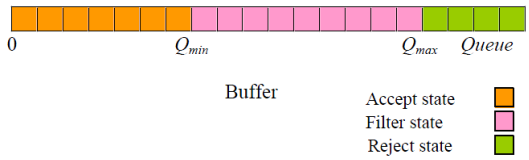


Figure 4. Buffer state

Every node which has data to send monitors its buffer and piggybacks its *WR* and *WQ* in its outgoing packets. If a node’s buffer occupancy exceeds a certain threshold and its data has higher priority among neighborhood, the corresponding congestion level bit in the outgoing packet header is set.

When congestion occurs, packets are dropped to alleviate congestion. Most of the queue schedulers drop packets from the tail rather than any position in the queue. But tail-dropping does not work well. For instance, if the queue in a sensor node is nearly full and dominated with low priority packets, when a high priority packet arrives, it is better to drop a low priority packet rather than the high priority packet. With tail-dropping, the high priority packet may be dropped due to queue overflow. Proposed protocol uses from this issue to mitigate the congestion.

In proposed protocol, the transmission is normal when the node state is ‘accept state’. In ‘filter state’, the packets are transmitted which the value of *WQ+WR* of its sender is high. In ‘reject state’, all received packets from all nodes are rejected and are routed to another path.

IV. Performance Evaluation

In this section, performances of the proposed protocol are shown by simulation. The scenario is similar to Fig. 1. The network stack of each node consists of IEEE 802.11 MAC layer. Simulation parameters are listed in Table I.

TABLE I. Simulation Parameters

Parameters	Value
Data rate	1 Mbps
Buffer size	30 packets
Number of sensors	20
Area	200 m *200 m
Q_{min}, Q_{max}	15, 25
Simulation time	100 sec
Data packet size	512 bits

Performance comparisons of the proposed protocol with PCCP protocol on throughput and fairness are provided as follows.

We study the impact of changing the source traffic rate on throughput. We change the traffic rate at the each source node from 20 to 100 packets/sec. We assume that priority of all nodes is same in this evaluation.

Normalized network throughputs of proposed protocol and PCCP are shown in Fig. 5.

As it can be seen, proposed protocol has performance better than PCCP in network throughput especially when that traffic rate at the source node is increased.

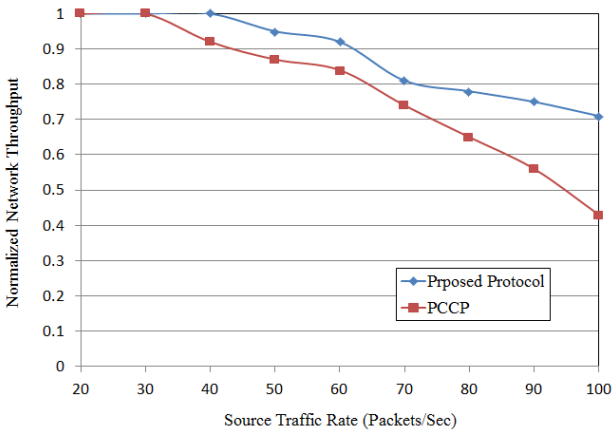


Figure 5. Normalized network throughput

V. Conclusion

In this paper, we propose a new protocol which can broadcast traffic on the entire network fairly. Proposed protocol adjusts every upstream traffic rate with its node dynamic priority to mitigate congestion. Simulation results show that the performance of proposed protocol is more efficient than previous algorithms especially in network throughput evaluation.

References

[1] Bashir Yahya, Jalel Ben-Othman, “Towards a classification of energy aware MAC protocols for wireless sensor networks”, Journal of Wireless Communications and Mobile Computing, Wiley.

[2] B. Hull, K. Jamieson, and H. Balakrishnan, “Mitigating congestion in wireless sensor networks, in Proc. ACM Sensys’04.

[3] Y. G. Iyer, S. Gandham, and S. Venkatesan, “STCP: A generic transport layer protocol for wireless sensor networks”, *in Proceedings of IEEE ICCCN* ,2005, Oct.17-19, San Diego, California, USA.

[4] Chonggang Wang, Kazem Sohraby, Victor Lawrence, Bo Li, Yueming Hu, “Priority-based Congestion Control in Wireless Sensor Networks”, IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing , Vol 1 (SUTC’06), 2006, pp. 22-31.

[5] C. T. Ee and R. Bajcsy, “Congestion control and fairness for many-to-one routing in sensor networks,” in Proc. Of ACM Sensys’04.

[6] B. Hull, K. Jamieson, and H. Balakrishnan, “Mitigating congestion in wireless sensor networks,” in Proc. ACM Sensys’04.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[7] C. Y. Wan, S. B. Eisenman, and A. T. Campbell, "CODA: Congestion detection and avoidance in sensor networks," in Proc. of ACM Sensys'03.

[8] A. Woo and D. C. Culler, "A transmission control scheme for media access in sensor networks," in Proc. Of ACM Mobicom'01.

[9] Yogesh S., O.B.Akan, Ian F.Akyildiz, "ESRT: Event-to-Sink Reliable Transport in Wireless Sensor Networks", in Proc. of Mobi- Hoc03, Annapolis, Maryland, USA, June, 2003.

[10] Chonggang Wang, Kazem Sohraby, Victor Lawrence, Bo Li, Yueming Hu, "Priority-based Congestion Control in Wireless Sensor Networks", IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing , Vol 1 (SUTC'06), 2006, pp. 22-31.

[11] B. Hull, K. Jamieson, and H. Balakrishnan, "Mitigating Congestion in Wireless Sensor Networks," in Proc. of ACM SenSys '04.

[12] Charalambos Sergiou, Vasos Vassiliou, Aristodemos Paphitis, 2013, Hierarchical Tree Alternative Path (HTAP) algorithm for congestion control in wireless sensor networks, Elsevier, Ad Hoc Networks 11 (2013) 257–272.

[13] Sajal K. Das, Chuang Lin, 2011, Traffic-Aware Dynamic Routing to Alleviate Congestion in Wireless Sensor Networks, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, 22(9), pp.1585-1599.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

EMAIL SECURITY USING CLUSTERING ALGORITHMS

Author:Tarushi Sharma

M-Tech(Information Technology)

Email-tarushisharma2323@gmail.com

Co-Author:Mrs.Amanpreet Kaur

(Assistant Professor)

Email-cecm.cse.akb@gmail.com

CGC Landran Mohali, Punjab,India

***Abstract*— Recent use of email analysis and data mining of email contents has proven to be useful in some sensitive places like national security agency to detect threats and fraud determination from terrorists. Moreover, it has been proved to be helpful for decision making, future team co-ordination, fraud detection and tracing the behavior of an employee. Using different clustering algorithms, we can find out similar patterns in emails for fraud detection. In this paper, we demonstrate how the popular k-means clustering algorithm can be profitably modified to make use of this information.**

***Index Terms*— Data mining, Email mining, Clustering algorithm, K-mean clustering algorithm, National Security Agency and Fraud Detection**

I. INTRODUCTION

In the research we have explained how to invoke filtering and learning schemes with the Explorer and link them mutually with the Knowledge Flow interface. To go in advancement, it is essential to learn somewhat about how Weka is put simultaneously.

Detailed, up-to-date information can be originated in the online credentials incorporated in the allocation. This is more technical than the descriptions of the learning and filtering schemes given by the more buttons in the Explorer and Knowledge Flow's object editors. In this paper, we initiate with a conversation of our infrastructure and then discuss our approach to mining the email archives; and in conclusion we present some preliminary consequences from our data analysis.

II. DATA MINING

Data mining- Data mining is the method of extracting useful information from the enormous quantity of data stored in the databases. Data mining tools and techniques help to foretell business trends those can happen in near potential. Association rule mining is an significant technique to determine hidden relationships between items in the contract. It is the computational process of discovering patterns in large data sets connecting methods at the meeting point of artificial intelligence, machine learning, statistics, and database system[2].

Data mining applications can use a range of parameters to inspect the data. They consist of association like patterns where one incident is associated to an additional event, such as purchasing a pen and purchasing paper, succession or path investigation like patterns where one incident leads to another incident, such as the birth of a child and purchasing diapers, sorting like detection of new patterns, such as coincidences among duct tape purchases and plastic canvas purchases, clustering like discovery and visually documenting groups of up to that time unknown facts, such as geographic position and brand preferences,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

and forecasting like discovering patterns from which one can formulate evenhanded predictions concerning potential actions, such as the calculation that people who connect an athletic club could take exercise classes.

a) Anomaly detection

Outlier/change/variation detection- The classification of extraordinary data reports, that might be motivating or data errors that necessitate further examination. [3]

b) Association rule learning –

Dependency modeling – Association rule mining is one of the major significant and well researched techniques of data mining for explanatory task, originally used for market basket investigation. It finds all the regulations existing in the transactional database that convince some minimum support and minimum assurance constraints. Classification using Association rule mining is an additional major Predictive analysis performance that aims to discover a small set of rule in the database that forms an precise classifier.

c) Clustering –

Clustering is a main task of explorative data mining, and a general technique for statistical data analysis used in numerous fields, counting machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. [7]

d) Classification –

It is the task of generalizing known configuration to apply to new data. For example, an e-mail program might challenge to classify an e-mail as "legitimate" or as "spam". Associative classification mining is a shows potential approach in data mining that utilizes the association rule detection techniques to construct categorization systems, also known as associative classifiers.

e) Regression –

Regression is a kind of arithmetical assessment technique used to map each data object to a real value provide calculation value Attempts to find a function which models the data with the smallest amount error. Uses of regression consist of curve fitting, prediction (counting forecasting), modeling of causal associations, and testing scientific hypotheses about relationships between variables.

III. EMAIL MINING

Email has met incredible reputation over the past few years. People are sending and receiving many messages per day, communicating with partners and friends, or exchanging files and information. Unfortunately, the occurrence of email overload has grown over the past years fetching a personal headache for users and a economic issue for companies. [5]

The paper will concern about how disciplines like Machine Learning and Data Mining can contribute to the resolution of the problem by constructing intelligent techniques which computerize email administration tasks and what advantages they hold over other predictable solutions. The particularity of email data and what special treatment it requires could be managed. Some interesting email mining applications like mail categorization, summarization, automatic answering and spam filtering will be also accessible [4]

The major purpose of email data mining is to present social network relationships and newly emerging parts of a social network. Due to the increasing threats to national security, people have started to use the results of email data mining to figure out terrorist threats. As of now, no one has tried to figure out future relations among the employees or even trace the behavior of an employee. Hence using email data mining and finding out future and behavioral relationships among users could

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

be extremely useful for an organization where the employee survey is often taking place. [11]

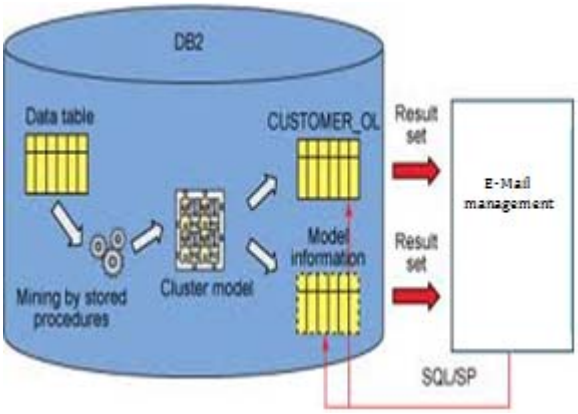


Fig 1: Email-Mining Process by taking an example [9]

In the above figure 1 a system of managing database issues like data collection, data storage, data mining are shown by taking an example of database of an organization. In some organizations, it is already in use to collect email statistical data and the progress of an employee compared to an employee sitting beside him. Many reputed companies are already in the process of using email log analysis to improvise spam detection, employee personalization and automated filling.

The manager of a particular team/group would be able to analyze the behavioral and social relationships between the users under his/her team, by just mining an email archive. As more and more research is being conducted in this area, some fruitful results have been produced in forensic analysis and national intelligence services. Last but not the least, in the area of decision making systems, researchers are more focused on email mining to achieve future decision making results for an organization. [8]

The impact of electronic mail in our daily life is now more obvious than ever. Each minute, millions of plain text or enriched messages are being sent and received around the globe. Some of them are read with extra care and, at the same time, many of them are deleted with obvious disinterest. As the internet grows, electronic mail has not only turned into a vital tool for our work but also into an important means of interpersonal communication. [10]

In professional life, email has invaded everywhere. Team organization, project management, information exchange Ducheneaut & Belloti, 2001, decision making, client support are only a few of a company's daily processes where email has been vital. Email also made personal communication significantly easier as it offered instant messaging with minimum cost. People from all over the world can now exchange opinions and information with such ease that made email the second most popular channel of communications after voice (Miller, 2003). [12]

Features that made email so popular are the rapidity of communication, the minimum cost and the fact that it is remarkably easy to use. An advantage over voice communication e.g. phone is that it is asynchronous, meaning that there is no need for both sides of communication to be on-line or in front of a computer at the same time.

Unfortunately, email could not escape the curse of Information Overload. Loads and loads of incoming messages have turned handling of electronic mail into a tedious task. Today, an average email user may receive at about 100 or 200 messages per day and, in a recent research, IDC1 predicts that by the year 2006 email traffic will be about 60 billion messages per day worldwide (Johnston, 2002).
Nowadays, people struggle to separate important messages that demand immediate attention from the mound and large companies are investing money in order to maintain email centers with personnel dedicated to answer client requests and queries sent by emails. Additionally, the problem of spam messaging has grown at a level that it is now considered an industry problem. It costs billions of dollars as it takes up bandwidth, clutters inboxes and occupies employees who are receiving them. Moreover, the content of many spam messages is unsuitable for children e.g. pornographic. [10]

IV. EMAIL LOG MINING

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The main purpose of email data mining is to present social network relationships and newly emerging parts of a social network. Due to the increasing threats to national security, people have started to use the results of email data mining to figure out terrorist threats. As of now, no one has tried to figure out future relations among the employees or even trace the behavior of an employee. Hence using email data mining and finding out future and behavioral relationships among users could be extremely useful for an organization where the employee survey is often taking place.

In some organizations, it is already in use to collect email statistical data and the progress of an employee compared to an employee sitting beside him. Many reputed companies are already in the process of using email log analysis to improvise spam detection, employee personalization and automated filling. [12]

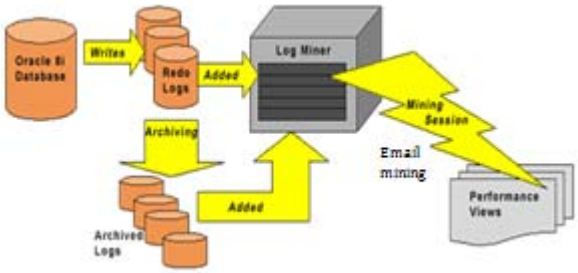


Fig. 2: Email Log Mining using a Log Miner

The manager of a particular team/group would be able to analyze the behavioral and social relationships between the users under his/her team, by just mining an email archive. As more and more research is being conducted in this area, some fruitful results have been produced in forensic analysis and national intelligence services. Last but not the least, in the area of decision making systems, researchers are more focused on email mining to achieve future decision making results for an organization.

Recent use of email analysis and data mining of email contents has proven to be useful in some sensitive places like national security agency to detect threats and fraud determination from terrorists. Moreover, it has been proved to be helpful for decision making, future team co-ordination, fraud detection and tracing the behavior of an employee.

In recent years, email has become a necessary part of any organization or group of the similar kind of users who share their information with each other on the network. Since email has become a necessary tool to communicate and co-ordinate with each other within an organization, the mining of email is sure to give some future decision and inter-relation based information. Let us consider an example of a software team within an organization. [13]

V. CLUSTERING ALGORITHM

Clustering can be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings including values such as the distance function to use, a density threshold or the number of expected clusters depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify pre-processing and parameters until the result achieves the desired properties. [11]

K-MEAN Algorithm

The k-means algorithm (Mac Queen 1967, Anderberg1973) is built upon four basic operations: selection of the initial k means for k clusters, calculation of the dissimilarity between an object and the mean of a cluster, allocation of an object to the cluster whose mean is nearest to the object, Re-calculation of the mean of a cluster from the objects allocated to it so that the intra cluster dissimilarity is minimized.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Except for the first operation, the other three operations are repeatedly performed in the algorithm until the algorithm converges. The essence of the algorithm is to minimize the cost function

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{i,l} d(X_i, Q_l)$$

where n is the number of objects in a data set **X**, X_i Î **X**, Q_l is the mean of cluster l, and y_i ,l is an element of a partition matrix $\mathbf{Y}_{nXl}$ as in (Hand 1981). d is a dissimilarity measure usually defined by the squared Euclidean distance. [12]

The biggest benefit of the *k*-means algorithm in data mining applications is its effectiveness in clustering bulky data sets. On the other hand, its use is inadequate to numeric values. The *k*-modes algorithm presented in this paper has removed this constraint at the same time as preserving its effectiveness.

The *k*-modes algorithm has prepared the subsequent extensions to the *k*-means algorithm:

1. Replacing means of clusters with modes,
2. Using new variation actions to deal with unconditional objects, and
3. Using a frequency based technique to update modes of clusters.

These extensions authorize us to use the *k*-means paradigm directly to cluster categorical data exclusive of need of data exchange.

Another benefit of the *k*-modes algorithm is that the modes give distinctive descriptions of clusters. These descriptions are very significant to the user in interpreting clustering consequences. Since data mining deals with very large data sets, scalability is a basic necessity to the data mining algorithms. Our investigational results have confirmed that the *k*-modes algorithm is indeed scalable to very large and composite data sets in terms of both the amount of records and the amount of clusters.

In fact the *k*-modes algorithm is more rapid than the *k*-means algorithm for the reason that our experiments have shown that the previous often needs less iteration to congregate than the later.

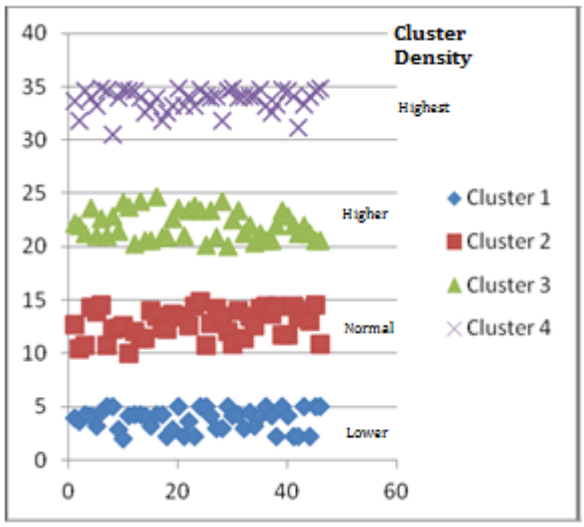


Fig. 3: Graph Shows Clusters in K-means Clustering Algorithm

There exist a few variants of the k-means algorithm which differ in selection of the initial k means, dissimilarity calculations and strategies to calculate cluster means (Anderberg 1973, Bobrowski and Bezdek 1991).The sophisticated variants of the k-means algorithm include the well-known ISODATA algorithm (Ball and Hall 1967) and the fuzzy k-means algorithms (Ruspini 1969, 1973).

VI. WEKA

In data mining, Weka is a program that evaluates the data analysis to develop models to support on decision making on business warehouse management. There are many data mining techniques for model developing. Among the most popular ones are Classification, Clustering and Association Rule Discovery which are applied in model developing Weka Machine Learning Project, 2010; Wass, 2007. For the models supporting decision making in business warehouse management

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

in this essay, Classification is used in analyzing nominal data and prediction analysis for numeric data.

A process in model developing with classification technique is applying training data in developing model and testing the model with the evaluation data. Then the model is used in real practice by applying the unseen data that there is still no answer class with this model, developed by Weka program.

In Weka, the implementation of a particular learning algorithm is encapsulated in a class. For example, the J48 class described previously builds a C4.5 decision tree. Each time the Java virtual machine executes J48, it creates an instance of this class by allocating memory for building and storing a decision tree classifier. The algorithm, the classifier it builds, and a procedure for outputting the classifier is all part of that instantiation of the J48 class. Larger programs are usually split into more than one class.

VII. CONCLUSION AND FUTURE WORK

The biggest advantage of the *k*-means algorithm in data mining applications is its efficiency in clustering large data sets. Using clustering algorithm we can find out similar patterns in any dataset. In emails we can find out similar pattern of text in emails of any region. And using K-mean clustering algorithm, that what is similarity between those patterns of email and how they are co-related to each other.

Our future work is to implement this technique using different statistical visualization tools and matching the similarity of patterns in text.

VIII. REFERENCES

[1] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic “"From Data Mining to Knowledge Discovery in Databases” Retrieved 17 December 2008.

[2]ACM SIGKDD "Data Mining Curriculum” 2006-04-30. Retrieved 2011-10-28

[3]Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic “"From Data Mining to Knowledge Discovery in Databases” Retrieved 17 December 2008.

[4]Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas” Email Mining: Emerging Techniques for Email Management”

[5] Email Mining: Emerging Techniques for Email Management Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas Aristotle University of Thessaloniki, Department of Informatics,

[6]. Hongjun Li, Jiangang Zhang, Haibo Wang and Shaoming Huang, "Mining Algorithm for Email's Relationships Based On Neural Networks". 2008 International Conference on Computer Science and Software Engineering

[7]. Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz and Anand Swaminathan, "Mining Email Social Networks". (May 22-23, 2006). Dept. of Computer Science, University of California, Davis.

[8]. R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. "Miningnewsgroups using networks arising from social Behavior". (2003).

[9]. Shlomo Hershkop, Ke Wang, Weijen Lee, Olivier Nimeskern, German Creamer, and Ryan Rowe, "Email Mining Toolkit Technical Manual". Department of Computer Science Columbia University, Version 3.6.8 - June 2006.

[10]. Bron, C. and J. Kerbosch. "Algorithm 457: Finding all cliques of an undirected graph." In Comm ACM, vol. 16, pg. 575-577, 1973.

[11]. Zuoliang Chen, Guoqing Chen BUILDING AN ASSOCIATIVE CLASSIFIER BASED ON FUZZY ASSOCIATION RULES International Journal of Computational Intelligence Systems, Vol.1, No. 3 (August, 2008), 262 – 273

[12]. Ranjana Vyas, Lokesh Kumar Sharma, Om Prakash vyas, Simon Scheider Associtive Classifiers for Predictive analytics: Comparative Performance Study, second UKSIM European Symposium on Computer Modeling and Simulation 2008.

[13].E. Ramaraj N. Venkatesan Positive and Negative Association Rule Analysis in Health Care Database, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.10, October 2008,325-330.

[14]. Khan, M.S. Muyeba, M. Coenen, F A Weighted Utility Framework for Mining Association Rules, Symposium Computer Modeling and Simulation, 2008. EMS '08. Second UKSIM European, page(s): 87-92.

[15] Wass, J. (2007). Weka Machine Learning Work bench Scientific Computing, 24(3), pp. 21-47.

[16] Mrs Amanpreet kaur (Assitant Professor), cecm.cse.akb@gmail.com.

I am Tarushi Sharma persuing M-Tech from CGC Landran,Mohali(2011-13) in Information Technology.I did my B-Tech from CCS University Campus,Meerut(2007-11) in Information Technology.

EMAIL SECURITY USING WEKA TOOL RESULTS OF K-MEAN CLUSTERING ALGORITHM.

Author:Tarushi Sharma

M-Tech(Information Technology)

Email- tarushisharma2323@gmail.com

Co-Author:Mrs.Amanpreet Kaur

(Assistant Professor)

Email-cecm.cse.akb@gmail.com

CGC Landran Mohali, Punjab,India

Abstract— Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Weka is a type of data mining tools. It is contain the many machine leaning algorithms. It provides the facility to classify our data through various algorithms. In this paper we are studying the various clustering algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters.

The K-means algorithm is a popular data-clustering algorithm. However, one of its drawbacks is the requirement for the number of clusters, K, to be specified before the algorithm is applied. This paper first reviews existing methods for selecting the number of clusters for the algorithm. Our main aim to show the comparison of the different samples of data like we are using different E-mails with similar text through different clustering algorithms of Weka and find out which parameter of Weka tool is effective for the users for data mining or email mining.

Keywords— Data mining algorithms, Weka tools, K-means algorithms, Clustering methods E-mail mining etc

I. INTRODUCTION

Data mining- Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

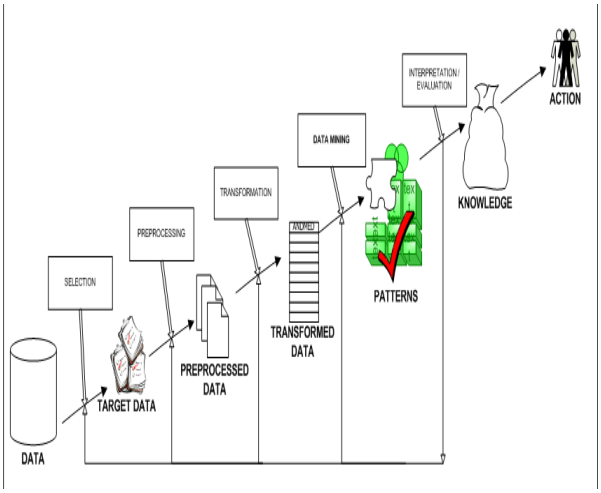


Figure1: Knowledge discovery in databases (KDD) and Data mining

II. CLUSTERING ALGORITHM

K-means algorithm- In centric-based clustering, clusters is represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the cluster centres and assign the objects to the nearest cluster centre, such that the squared distances from the cluster are minimized.

The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximate method is Lloyd's algorithm, [3] often actually referred to as "k-means algorithm". It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centric to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means).

Most k-means-type algorithms require the number of clusters -- to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centric. This often leads to incorrectly cut borders in between of clusters (which are not surprising, as the algorithm optimized cluster centers, not cluster borders).

K-means has a number of interesting theoretical properties. On one hand, it partitions the data space into a structure known as Voronoi diagram. On the other hand, it is conceptually close to nearest neighbor classification and as such popular in machine learning.

Weka tool- Data mining [16] isn't solely the domain of big companies and expensive software. In fact, there's a piece of software that does almost all the same things as these expensive pieces of software the software is called WEKA. WEKA is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. It uses the GNU General Public License (GPL). The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results (think tables and curves). It also has a general API, so you can embed WEKA, like any other library, in our own applications to such things as automated server-side data mining tasks.

In this paper weka is used to analyse and experiment the email management using the all clustering algorithms of weka. To complete to this purpose I am taking data as different email with similar text and compare and analyze them using weka clustering algorithm. And study all clustering [5] algorithms of weka classification of data. For working of weka we does not need the deep knowledge of data mining that’s reason it is very popular data mining

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

tool. Weka also provides the graphical user interface of the user and provides many facilities [4, 7].

III. WEKA STRUCTURE

We have explained how to invoke filtering and learning schemes with the Explorer and connect them together with the Knowledge Flow interface. To go further, it is necessary to learn something about how Weka is put together. Detailed, up-to-date information can be found in the online documentation included in the distribution. This is more technical than the descriptions of the learning and filtering schemes given by the more buttons in the Explorer and Knowledge Flow's object editors. It is generated directly from comments in the source code using Sun's Javadoc utility. To understand its structure, you need to know how Java programs are organized.

Classes, instances, and packages- Every Java program is implemented as a class. In object-oriented programming, a class is a collection of variables along with some methods that operate on them. Together, they define the behaviour of an object belonging to the class. An object is simply an instantiation of the class that has values assigned to all the class's variables. In Java, an object is also called an instance of the class.

The weka.classifiers package- The classifiers package contains implementations of most of the algorithms for classification and numeric prediction described in this book. The most important class in this package is Classifier, which defines the general structure of any scheme for classification or numeric prediction. Classifier contains three methods, buildClassifier(), classifyInstance(), and distributionForInstance().

IV. EMAIL MINING THROUGH WEKA

Clustering of spam messages means automatic grouping of thematically close spam messages. In case of information streams as E-mails, this problem becomes complicated necessity to carry out this process in real-time mode. There are some complications connected with plurality of a choice of algorithms for clustering of spam messages. Different methodologies use different similarity algorithms for electronic documents in case of a considerable quantity of signs.

As soon as classes are defined by clustering method, there is a necessity of their support as spam constantly changes, and spam messages collection replenishes. In considered work, the new algorithm for definition of criterion function of spam messages clustering problem is offered. The clustering problem itself is solved by genetic algorithm [28]. Weka tools are the subjects of many scientific works.

WEKA- In email mining, Weka is a program that evaluates the data analysis to develop models to support on decision making on business warehouse management. There are many data mining techniques for model developing. Among the most popular ones are Classification, Clustering and Association Rule Discovery which are applied in model developing (Weka Machine Learning Project, 2010; Wass, 2007). For the models supporting decision making in business warehouse management in this essay, Classification is used in analyzing nominal data and prediction analysis for numeric data.

V. IMPLEMENTATION OF WEKA TOOL

My methodology is very simple. I am taking the past project data from the repositories and apply it on the weka. Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The k-means algorithm is best suited for implementing this operation because of its efficiency in clustering large data sets.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

We introduce new dissimilarity measures to deal with categorical objects, replace means of clusters with modes, and use a frequency based method to update modes in the clustering process to minimise the clustering cost function.

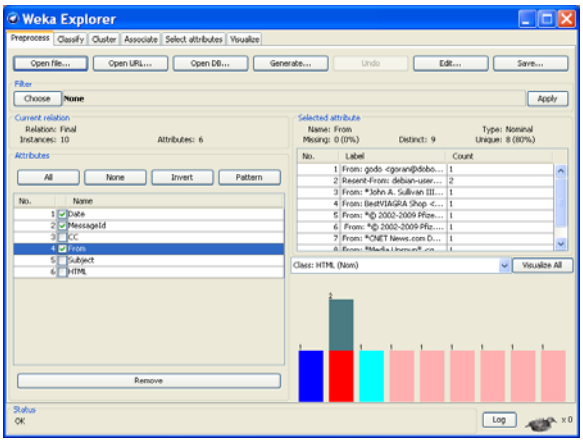


Figure 2: Weka explorer showing different data clusters

However, working only on numeric values limits its use in data mining because data sets in data mining often contain categorical values. In this paper we present an algorithm, called k-modes, to extend the k-means paradigm to categorical domains.

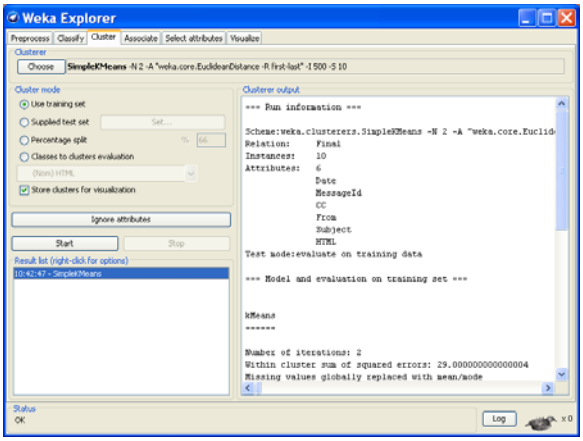


Figure 3: Clustering algorithm implementation

This paper focuses on the classification of textual spam E-mails using data mining techniques. Our purpose is not only to filter messages into spam and not spam, but still to divide spam messages into thematically similar groups and to analyze them, in order to define the social networks of spammers.

VI. WEKA RESULTS

In the below figure we have different types of email cluster with their unique message id.

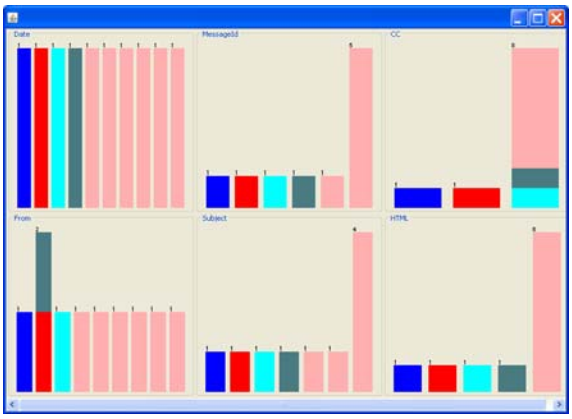


Figure 4: Email classification through Weka

Weka uses different clustering algorithm to make cluster of similar text messages and the k-means algorithm is the simplest of them. Above figure shows result from k-nodes cluster and have differentiate among important email to spam.

VII. CONCLUSION

A new idea that seems promising is Semantic Email. In parallel with Semantic Web, Email could be enriched with meta-tags in order to describe better the information included in the message. As discussed in (McDowell, Etzioni, Halevy, & Levy, 2004), applications like Information Dissemination, Event Planning, Report Generation and Email Auctions / Giveaways could be achieved with the use of the Semantic Email.

Email mining raised new difficulties and challenges for the text mining community. New solutions had to be proposed in already discussed areas due to email data peculiarity. Additionally, domain-specific problems provoked the development of new applications like spam filtering, email answering and thread summarization.

While effective solutions have been proposed to most email problems, not all of them have been implemented in popular email clients. In fact, with the exception of spam filtering which is now integrated in most commercial clients, no other applications have been used widely from the average user. There is therefore an obvious need to implement those methods and integrate them into useful and accurate software which will let people take back control of their mailboxes.

VIII. REFERENCES

[1]R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In WWW '03: Proceedings of the 12th international conference on World Wide Web, 2003.

[2] A.-L. Barab´asi and R. Albert. Emergence of scaling in random networks. Science, 286:509–512, 1999.

[3] C. Bird, A. Gourley, P. Devanbu, A. Swaminathan, and M. Gertz. Mining email social networks in postgres. In MSR '06: Proceedings of the International Workshop on Mining Software Repositories, 2006.

[4] F. Brooks. The Mythical Man-Month: Essays on Software Engineering, 20th Anniversary Edition. Addison-Wesley, 1995.

[5] S. Chapman. Sam's string metrics page. www.dcs.shef.ac.uk/sam/stringmetrics.html.

[6] J. F. P. D. Cleidson de Souza. Seeking the source: Software source code as a social and technical artifact, 2005. http://opensource.mit.edu/papers/desouza.pdf.

[7] K. Crowston and J. Howison. The social structure of free and open source software development. opensource.mit.edu/papers/crowstonhowison.pdf, November 2004.

[8] B. J. Dempsey, D. Weiss, P. Jones, and J. Greenberg. Who is an open source software developer? Communications of the ACM, 45(2):67–72, February 2002.

[9] L. C. Freeman. Centrality in social networks I. Conceptual clarification. Social Networks, 1:215–239, 1979.

[10] Segal, R., & Kephart, J. O. (2000). *Incremental Learning in SwiftFile.* Paper presented at the 7th International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA.

[11] Strehl, A., Ghosh, J., & Mooney, R. (2000). *Impact of Similarity Measures on Web-page Clustering.* Paper presented at the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI 2000), Austin, Texas, USA.

[12] Surendran, A. C., Platt, J. C., & Renshaw, E. (2005). *Automatic Discovery of Personal Topics to Organize Email.* Paper presented at the 2nd Conference on Email ant Anti-Spam, Stanford University, USA.

[13] Vel, O. d., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *SIGMOD Rec., 30*(4), 55-64.

[14] Wan, S., & McKeown, K. (2004). *Generating Overview Summaries of Ongoing Email Thread Discussions.* Paper presented at the COLING 2004, the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

[15] Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

[16] Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization.* Paper

presented at the ICML-97, 14th International Conference on Machine Learning, Nashville, US.

[17] Zechner, K. (1996). *Fast generation of abstracts from general domain text corpora by extracting relevant sentences.* Paper presented at the 16th conference on Computational Linguistics, Copenhagen, Denmark.

[18] Neto, J. L., Freitas, A. A., & Kaestner, C. A. A. (2002). *Automatic Text Summarization Using a Machine Learning Approach.* Paper presented at the SBIA '02: 16th Brazilian Symposium on Artificial Intelligence

[19] Pazzani, M. J. (2002). *Representation of electronic mail filtering profiles: a user study.* Paper presented at the Intelligent User Interfaces.

[20] Platt, J. C. (1999). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Scholkopf, C. Burges & A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning* (pp. 185-207): MIT Press.

[21] Porter, M. F. (1997). An algorithm for suffix stripping. In *Readings in information retrieval* (pp. 313-316): Morgan Kaufmann Publishers Inc.

[22] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*, 81-106.

[23] Mrs Amanpreet kaur (Assitant Professor), cecm.cse.akb@gmail.com.

I am Tarushi Sharma persuing M-Tech from CGC Landran,Mohali(2011-13) in Information Technology.I did my B-Tech from CCS University Campus,Meerut(2007-11) in Information Technology.

IJCSIS AUTHORS' & REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Mrs Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Mr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Mr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Mr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Mr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Mr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Mr. P. Vasant, University Technology Petronas, Malaysia
Mr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Mr. Praveen Ranjan Srivastava, BITS PILANI, India
Mr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Mr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Mr. Tirthankar Gayen, IIT Kharagpur, India
Ms. Huei-Ru Tseng, National Chiao Tung University, Taiwan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2014
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/